In the Name of God

Journal of
Information Systems & Telecommunication
Vol. 4, No. 3, July-September 2016, Serial Number 15

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology

Affiliated to: Academic Center for Education, Culture and Research (ACECR)

Managing Director: Habibollah Asghari, Assistant Professor, ACECR, Iran
Editor in Chief: Masoud Shafiee, Professor, Amir Kabir University of Technology, Iran

Administrative Manager: Shirin Gilaki
Executive Assistant: Behnoosh Karimi
Art Designer: Amir Azadi
Print ISSN: 2322-1437
Online ISSN: 2345-2773
Publication License: 91/13216

Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,
P.O.Box: 13145-799 Tel: (+9821) 88930150 Fax: (+9821) 88930157
Email: info@jist.ir
URL: www.jist.ir

Indexed in:

- Index Copernicus International	www.indexcopernicus.com
- Journal of Information Systems and Telecommunication	www.jist.ir
- Islamic World Science Citation Center (ISC)	www.isc.gov.ir
- Scientific Information Database (SID)	www.sid.ir
- Regional Information Center for Science and Technology (RICeST)	www.ricest.ac.ir
- Magiran	www.magiran.com

Publisher:
Regional Information Center for Science and Technology **(RICeST)**
Islamic World Science Citation Center **(ISC)**

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Digital & Signal Processing Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

- Abdolvand Neda, Alzahra University, Tehran, Iran
- Ahari Lahegh Alireza, Islamic Azad University, Qazvin Branch, Qazvin, Iran
- Akbarian Behnam, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- AleAhmad Abolfazl, University of Tehran, Tehran, Iran
- Buyer Asgar Ali, Azarbijan Shahid Madani University, Tabriz, Iran
- Ebrahimzadeh Ataollah, Babol Noshirvani University of Technology, Babol, Iran
- Ebrahimi Moghadam Abbas, Ferdowsi University of Mashhad, Mashhad, Iran
- Emami Hossein, Islamic Azad University, Majlesi Branch, Isfahan, Iran
- Fadaei Eslam Mohamamd Javad, Semnan University, Semnan, Iran
- Faghihi Mehdi, Islamic parliament Research Center, Tehran, Iran
- Falahati Abolfazl, Iran University of Science and Technology, Tehran, Iran
- Fallah Mohammad, Islamic Azad University, South Tehran Branch, Tehran, Iran
- Fathi Abdolhossein, Razi University, Kermanshah, Iran
- Fotouhi Faranak, Qom University, Qom, Iran
- Ghaffari Ali, Islamic Azad University, Tabriz Branch, Tabriz, Iran
- Ghanei Yakhdan Hossein, Yazd University, Yazd, Iran
- Ghanbari Mohammad, University of Essex, Colchester, UK
- Ghasemzadeh Mohammad, Yazd University, Yazd, Iran
- Hamidi Hojjatollah, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Hashemi Sattar, Shiraz University, Shiraz, Iran
- Hosseini Monireh, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Jamali Shahram, University of Mohaghegh Ardabili, Ardabil, Iran
- Jannati Hoda, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
- Maham Behrouz, University of Tehran, Tehran, Iran
- Mahdavian Hojjat, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Marvi Hossein, Shahrood university of Technology, Shahrood, Iran
- Mirian Hosseinabadi Seyed Hasan, Sharif University of Technology, Tehran, Iran
- Mirroshandel Seyed Abolghasem, University of Guilan, Rasht, Iran
- Mohammad Khanli Leili, University of Tabriz, Tabriz, Iran
- Nasri Mehdi, Shahid Bahonar University of Kerman, Kerman, Iran
- Oveisi Farid, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Rafe Vahid, Arak University, Arak, Iran
- Rahmani Manesh Mohammad, Semnan University, Semnan, Iran
- Rasi Habib, Shiraz University of Technology, Shiraz, Iran
- Sharifi Alireza, University of Tehran, Tehran, Iran
- Siami Mohammad, Iran University Of Science and Technology, Tehran, Iran

Table of Contents

A Bio-Inspired Self-configuring Observer/ Controller for Organic Computing Systems

Ali Tarihi
Department of Computer Engineering and Science, Shahid Beheshti University, Tehran, Iran
a_tarihi@sbu.ac.ir
Hassan Haghighi*
Department of Computer Engineering and Science, Shahid Beheshti University, Tehran, Iran
h_haghighi@sbu.ac.ir
Fereidoon Shams Aliee
Department of Computer Engineering and Science, Shahid Beheshti University, Tehran, Iran
f_shams@sbu.ac.ir

Received: 04/Jul/2015 Revised: 29/Jun/2016 Accepted: 25/Jul/2016

Abstract

The increase in the complexity of computer systems has led to a vision of systems that can react and adapt to changes. Organic computing is a bio-inspired computing paradigm that applies ideas from nature as solutions to such concerns. This bio-inspiration leads to the emergence of life-like properties, called self-* in general which suits them well for pervasive computing. Achievement of these properties in organic computing systems is closely related to a proposed general feedback architecture, called the observer/controller architecture, which supports the mentioned properties through interacting with the system components and keeping their behavior under control. As one of these properties, self-configuration is desirable in the application of organic computing systems as it enables by enabling the adaptation to environmental changes. However, the adaptation in the level of architecture itself has not yet been studied in the literature of organic computing systems. This limits the achievable level of adaptation. In this paper, a self-configuring observer/controller architecture is presented that takes the self-configuration to the architecture level. It enables the system to choose the proper architecture from a variety of possible observer/controller variants available for a specific environment. The validity of the proposed architecture is formally demonstrated. We also show the applicability of this architecture through a known case study.

Keywords: Organic Computing; Observer/ Controller Architecture; Self-* Properties; Self-Configuration; Formal Verification.

1. Introduction

The arising complexity in computer systems has led to the introduction of new paradigms such as Autonomic Computing [1], Organic Computing (OC) and Pervasive Computing [2] that cope with complexity. Organic Computing is centered around cooperating entities which are sometimes called agents [2]; each of which has a set of capabilities. These capabilities are mostly sensors and actuators that enable the agents to interact with their environment and perform what is expected from them. Agents are also capable of communicating with each other and ultimately contribute to the creation of a single collective OC system. Because of the complexity in OC systems, an explicit design cannot be given for each possible situation. Therefore, a degree of freedom in decision making is given to the agents, so that the system can be managed collectively based on the local decisions [3]. This leads to the emergence of properties, like self-healing, self-configuration, and self-optimization at the system level that are called self-* in general [2].

The main drawback of obtaining self-* properties in this manner is the possible emergence of unwanted behaviors due to the lack of system-wide vision in the local decisions. Coping with this problem implies using a control mechanism. To achieve this goal, the Observer/Controller architecture or o/c (for short) has been proposed for OC systems [3]. The observer as the name suggests has to observe the system passively and reports to the controller for proper actions. The part of the system that is under observation is usually called System under Observation and Control (SuOC) [2]. The generic o/c architecture [3] is the most known and cited o/c architecture, and many of the existing researches in OC refine the generic o/c architecture for their own purposes; for example, see [4]-[6]. The generic o/c architecture is studied deeper in Section 2.

Self-configuration, which is related to the ability of the system to reconfigure itself dynamically [7], is among self-* properties that the o/c architecture tries to control. It is defined as “the set of all system and environmental attributes that can be modified by control actions” [8]; these attributes are divided into two categories: internal and external [8]. The former attributes that controlled by the system while the latter are controlled by the user or an external entity.

* Corresponding Author

In OC, self-configuration is mainly achieved in the SuOC level, meaning that the SuOC is reconfigured accordingly by the o/c component. In this way, the benefit of the self-configuration property is not present in the component level, or in other words, OC systems are committed to have a fixed o/c component governing the SuOC. Therefore, any rearrangement or change in the o/c component is prevented, which will be a major drawback to environments where multiple o/c components configurations are applicable. This issue motivated us to enable the self-configuration property at the o/c component level and achieve a first step toward a self-* enabled o/c component for the o/c architecture.

Hence, the main contribution of this paper is focused on promoting the self-configuration property to the o/c component level. In order to achieve this goal, we propose a bio-inspired self-configuring o/c architecture that configures itself according to the operational parameters. The bio-inspiration in our work comes from the notion of cell differentiation process [9]. We also use the feature model concept from the software architecture, or more precisely, the software product line so as to capture o/c component configurations. In addition, we present and evaluate our ideas using formal methods.

For this purpose Section 2 is dedicated to the background concepts, especially the biological ones while Section 3, the related work is reviewed, and then the proposed o/c architecture is presented in Section 4. In order to validate the proposed architecture, it is specified and verified using formal methods in Section 5. Section 6 shows the applicability of the proposed o/c architecture through a known case study, and finally, the last section is devoted to the conclusion and some directions for future work.

2. Background

2.1 The Generic o/c Architecture

The generic o/c architecture [3] consists of a set of components shown in Fig. 1. The observer component in this architecture is composed of several sub-components that monitor and use data from the SuOC for analysis and prediction; the results are aggregated and then used by the controller.

The controller component is in charge of executing the decisions made by its learning components for the SuOC. Three sources of data are given to the “aggregator”, and then the aggregated data is used by the “mapping” (rule base) and “rule performance evaluation” subcomponents of the controller. This component has both online and offline learning subcomponents. The “rule performance evaluation” subcomponent is for online learning, which updates the existing rules as needed, whereas the “rule adaptation” and “simulation model” subcomponents are due to offline learning, they create new rules and delete the old ones. The “objective function” represents the user interactions that affect the control of the system. Finally, an “observation model”, which is applied by all the observers in the OC system, is selected by the controller to indicate the observable attributes and the proper analysis method and parameters for observation (e.g., the sampling rate).

The generic o/c architecture has three variants [3]: The centralized variant that consists of a single o/c component and a single SuOC, while decentralized variant has many SuOCs, each with a dedicated o/c component. The third variant is the multi-level o/c architecture, in which one of the o/c components is in the highest level, while the underlying SuOC consists of a collection of smaller SuOCs. These smaller SuOCs in turn can have their own SuOCs, resulting in a fractal like structure.

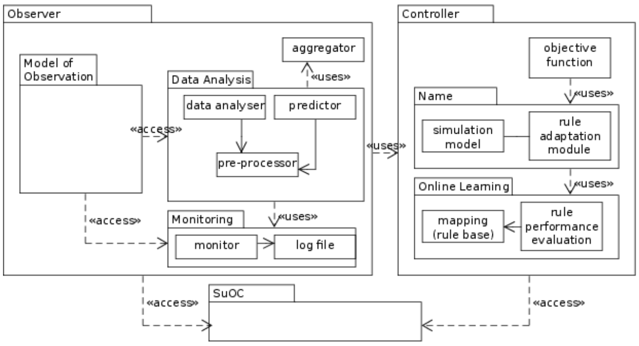


Fig. 1. The generic o/c architecture [3]

2.2 Cell Differentiation

For introducing the notion of cell differentiation, it is helpful to have a few words about the functionality of cells and the difference between them.

Multicellular organisms (or metazoons) need different types of cells (e.g., blood cells and neurons) so as to survive. Each cell type has a variety of functions to perform, some are common to all, while others are special to that type of cell. From a Biochemical point of view, proteins contribute to any biological function, and therefore, the difference between the cells comes from the difference in the proteins they have. For example, red blood cells have the special function of transferring oxygen in the blood because of hemoglobin, a protein that they have.

Regarding the mentioned concepts, interesting questions arise: 1) where proteins come from and 2) what makes each cell produce a special subset of proteins? The precise answer to this question is a major research topic in modern biology. But, we intend to present a brief answer from the biology literature that is both related and useful for the bio-inspiration mechanism used in this paper. All proteins inside a cell are encoded in a large biochemical molecule named DNA, which has many sections called genes that are used in a process called “transcription” [9]. In this process, the cell produces proteins from the DNA (the answer to the first question).

When a gene is used in creating proteins, it is said to be expressed. The term ``repressed'' is employed when a gene is not used for some reason such as some chemicals [9]. In other words, the expression/repression of genes controls the function of cells via proteins. This means, that the difference in the proteins produced by the cells comes from the expression/repression of their genes (the answer to the second question).

With this introduction, cell differentiation can be defined as follows. All of the multicellular organisms begin in an embryonic state (before the birth) from a single cell called zygote, and all the cells evolve from it.

With each generation, some genes are expressed/repressed, and ultimately, specialized cells are evolved. This process which is most active before the birth is known as cell differentiation, which has critical role in the life of multicellular organisms. There are some decisive factors that affect cell differentiation [9], especially the gene expression/repression, that results in different functionalities in the cell. Cell differentiation also depends on some chemicals, like growth factors and inducers, which can cause or prevent cell differentiation [9]. Another factor that affects cell differentiation is the micro-environment (also called niche) which surrounds the cells. For instance, keratinocytes (skin cells) are affected by the micro-environment, and in this way, specialize and form the skin [9].

This is only a brief introduction to cell differentiation, the interested reader is referred to [9] for more information.

2.3 Feature Model

The feature model comes from Feature-Oriented Domain Analysis [10] “describe a hierarchy of properties of domain concepts” [11]. This model helps to determine which combinations of features can be selected for domain concepts. If we consider the domain of wrist watches as an example, some of the general statements that can be given are: The watch can be either digital or mechanical, displaying the time by digits or hands. , and in some showing the date.

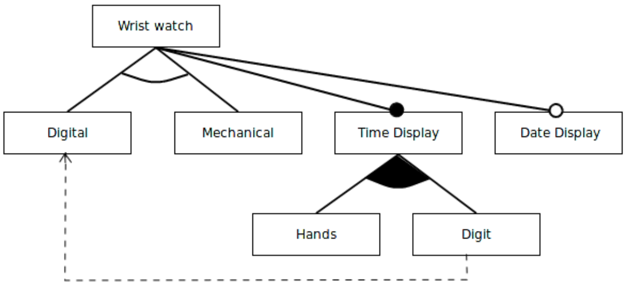


Fig. 2. A simplified feature diagram of the wrist watch example [10].

Feature diagram is the graphical representation of a feature model. Fig. 2 is a simplified feature diagram of the wrist watch example. The full dots indicate the “mandatory” features (like Time Display) that must be present in any domain concept regarding this feature model, while the empty dots indicate an “optional” property (like Date Display). The arc between Digital and Mechanical denotes “alternative” relationship (i.e., only one of these two features must be selected). There is another “or-relation” (for example, between Hands and Digit) that indicates any number of features that can exist together (e.g., a digital watch can have either digital hands or digits). Other two common relationships are “require” and “exclude” relationships [12]. The “require” relationship between Digit and Digital represented by a dashed arrow indicates that Digit display cannot be selected without a digital wrist watch. For example, digit display is only available to digital wrist watches. The “exclude” relationship is used to indicate that two features cannot exist together. It is usually displayed by double headed dashed arrows in feature diagrams. Having these relationships, feature models have many uses. In software product lines they are used for defining products and configurations [11]. In Section 4, we use the feature models for the configuration definition.

3. Related Work

Brinkschulte et al. [13] proposed an OC operational mechanism called Artificial Hormone System for task distribution among heterogeneous processing elements based on three types of hormones, namely, eager value, suppressor, and accelerator. The eager value determined the appropriateness of a task to be executed on a processing element. The suppressor and accelerator had two opposite effects on the process elements. The former increased the chance for taking tasks, while the latter tried to repress the execution of tasks. The Artificial Hormone System achieve

d many self-* properties by employing various sub-types of these three hormones [2] that participate in a hormone based control loop [2], in which each process element declares the appropriateness of a given task execution. Hormones from other process elements affected appropriateness value declared by the process elements. The overall effects of hormones on the control loop decided which process element would execute the task. The OC system achieved self-configuration by finding a suitable initial configuration for tasks based on the function of these hormones.

Roth et al. [14] suggest an OC middleware consisting of an organic manager and a set of ordinary services (like a database service) that communicated via the middleware running on distributed nodes for ubiquitous and pervasive computing. The goal of the middleware was to enable self-* properties (including self-configuration) for ordinary services. In this regard, the organic manager monitored the middleware and incorporated some self-* services, each of which was responsible for one self-* property. Using these self-* services required specific information provided by each ordinary service. However, since self-* services are independent, they might make conflicting decisions. Using the approach of Satzger et al. [5], a high-level planner component was added to the middleware in order to resolve the possible conflicts.

The o/c component of the middleware was inspired from the MAPE cycle of IBM autonomic computing [1] consisting of “Monitor”, “Analysis”, “Plan” and “Execute” stages. In the monitor stage, an information pool manager component managed the information pools containing the information needed for the control mechanism. The analyze stage had an event manager and a fact base components; when an event occurred, in order to use the event for planning, the event was transformed into facts. The plan stage, consisted of both a low level planner and a high level planner components. A plan was devised and then executed so as to solve any detected problem using these two planners. The low level planner component had a reflex manager component that managed the low level reflexes subcomponent. The reflexes subcomponent acted like a cache for previous system rules. Having this cache, if a previous decision was applicable to the current state, it would be applied. The high-level planner finds solutions to situations that are not solved by the low level planner. The high-level planner is managed by the high-level manager that converts the facts into a high-level language to solve by the planner. Finally, the actuator executes the plan given by the plan stage. Using this structure, the self-configuration service in this middleware determines the required resources for ordinary services and "triggers an auction" [14] so as to find the best node for that service.

Nafz et al. [6] proposed the Restore Invariant Approach (RIA) controller in which a set of reconfiguration algorithms processed a set of resources and agents having the required capabilities. The OC system tried to keep a set of invariants, regarding these invariants, result checker component examined the results of the used reconfiguration algorithms before the actual reconfiguration. Reconfiguration algorithms component was responsible for achieving self-configuration and was used in determining which capability must be active on which agent (as the initial configuration). These algorithms were also used for reconfiguring the agents whenever the invariant was violated.

The ORCA project was aimed at "transferring self-* properties to robotic systems" [15]. In this project a multi-level o/c architecture with decentralized modules was proposed. One type of these modules included Organic Control Units that monitored and controlled other modules and configured them for operation. The lower-level organic control units were themselves monitored by higher-level organic control units leading to a multi-level self-configuration mechanism.

In summary, it can be said that all of the mentioned works only covered self-configuration in the SuOC level and do not extend it to the o/c component; hence, it lost the advantages of self-configuration in this level by having a fixed o/c component. The fixed architecture prevents any rearrangement or change in the o/c components, which will be a major drawback to environments where multiple o/c component configurations are applicable.

4. Proposed Architecture

Our approach to enabling bio-inspired self-configuring o/c (sco/c) is architectural. We try in this section, which is divided into several subsections, to explain the rationale behind our architectural decisions. First, we explain the influence of the bio-inspiration from the cell differentiation on our architectural decisions as principles extracted from cell differentiation. Then, an illustrative example is introduced that will be used throughout the paper for demonstrating our proposed architecture. The third subsection presents an architectural meta-model that incorporates our core ideas.

4.1 Bio-inspiration for Self-Configuration

The cell differentiation process can be considered as an advanced form of self-configuration in which each cell self-configures its functionality accordingly. To be able to apply the benefits of cell-differentiation, we need to have building blocks analogous to the cells. This leads us to the agents and the first core principle in sco/c architecture.

Principle 1. In sco/c architecture, the system is considered as a collection of communicating agents.

Though this principle is not novel, it is required as a base for the application of the other bio-inspired principles. Based upon Principle 1, we can adopt the concept of genes. The difference between the cells is related to the expressed/repressed genes. This must be shown in the sco/c architecture, too. Subsequently, we must be able to express the system in terms of genes, which their active/inactive state affects the behavior of the agents and ultimately the system.

Just like the multicellular organisms, where everything is expressed through the genes, we need an alternative concept so as to capture the sco/c architecture. We propose the use of the "capability" concept that has also been used in organic computing [2] as well as multi-agent systems.

Principle 2. For all the agents, every function must be definable in terms of capabilities. Every functionality is available if and only if the corresponding capability is activated. Likewise, the deactivation of any capability will result in the lack of corresponding functionality.

This is analogous to gene expression/repression in cells. This principle shows what the agents do. In relation to this principle the question of that what should be done about the "capabilities" of the o/c component may arise, which can be answered in various ways. Regarding the bio-inspiration, it can be noticed that all the functionalities of a living organism, even the control mechanisms, are coded into the genes. Since we have chosen capabilities as counterpart of the genes, the control mechanism of the system must be represented in terms of capabilities.

This is a key principle in sco/c architecture that results in a uniform view of the system that makes the agents more like cells in multicellular organisms. This means everything, including the control mechanism is represented using one concept. This principle blurs the distinction between SuOC and the o/c component compared to other o/c architectures. So as to simplify the architecture description, we distinguish the capabilities representing the o/c component from the rest of the capabilities.

We promote the concept of agent capabilities by introducing another set of capabilities called Organic Computing capabilities.

Principle 3. The Organic Computing capabilities or OC capabilities are related to the o/c component. They participate in the observation and control of the OC system. The set of OC capabilities includes the sub-components of the o/c component and follow Principle 2 in terms of activation and deactivation.

In order to distinguish between the OC capabilities and the capabilities that have nothing to do with the control mechanism, we will refer to the latter as normal capabilities. In other words, agents use normal capabilities in performing their normal tasks. This includes the agent sensors and actuators for interacting with their environment.

For example, when the RIA controller is identified as the suitable o/c component, the invariant monitor, reconfiguration algorithms and result checker are the needed OC capabilities. In addition, the "reconfiguration algorithms" capability needs the "invariant monitor" capability, while in turn it is needed for the "result checker" capability.

Principle 4. In order to form the control mechanism, the required relationships between the OC capabilities must be established.

For example, an OC capability like "data analyzer" from the generic o/c architecture (Section 2), so as to operate, needs to be somehow connected to a monitoring OC capability. In this way, a set of relationships between the OC capabilities is formed. It can be said that o/c architecture can be realized via cooperation of agents using OC capabilities with regard to their relationships. So far, the presented principles can create the foundation needed for sco/c architecture. The self-configuration property of the sco/c architecture is also influenced by bio-inspiration as follows. In the beginning stages of cell differentiation, only zygote exists with no differentiation. After that, some genes are expressed in the following generations, and thus, specialized cells appear.

Principle 5. In the beginning, no OC capability is "active"

The control mechanism is the first thing to be realized. Since the control mechanism is realized by OC capabilities, OC capabilities must be activated in such a way that the relationships between them are preserved.

This principle ensures that the system only operates when there is a control mechanism formed using OC capabilities (Principle 4). This principle prevents the system from operating without a control mechanism.

Principle 6. Micro-environment and the chemicals present in it are required for the cell differentiation process.

The micro-environment is achieved using the concept of neighborhood that is common in multi-agent systems meaning that when an OC capability is active in a neighborhood, it can prevent the other agents from activating it. Also, when a needed OC capability is absent from a neighborhood, it must be activated. For the chemicals (for example, inducers and growth factors), messaging will be used. Similarly, when an OC capability residing in a different agent is needed by another OC capability (having "require" relationship) messaging is used.

Principle 7. Each cell differentiates using its genes. Gene expression/repression play a key role in deciding what gene should be expressed/repressed.

This principle implies that local control of gene expression/repression is needed. Each agent must know the relationship between the OC capabilities and when it should activate them, and it must be able to activate/deactivate them when needed.

Based on these principles, the sco/c architecture can be presented, but first, an illustrative example is presented in the next subsection in order to help understand the application of the bio-inspired principles in the sco/c architecture.

4.2 Illustrative Example

The example is a self-organizing resource-flow system [5], [6] and [16] in which a number of resources are processed by independent agents. The process of each resource consists of a set of tasks performed on each resource by the agents. Each agent has a collection of tools, each of which can perform a specific task. These tools might fail, rendering the agent unable to perform one or more of its tasks. The goal is to reconfigure the agents in a way that the processing of resources can still continue. The reconfiguration mechanism changes the assignment of tools to the agents, or in other words, changes the tasks they perform. It must be noted that at some point no reconfiguration can be done so as to keep the process going on. For instance, when all the instances of a tool is broken, no agent can perform the task related to that tool anymore. This will leave no possible reconfiguration. The number of tasks for the resources is not restricted to any specific number, but in [5] , [6] and [16] three tasks for each resource were considered for identical agents, which were drilling a hole in the resource (it a work piece), inserting a screw in it and tightening the screw.

In order to keep the illustrative example simple and tangible as possible, we will use this particular instance of self-organizing resource-flow system (as defined in Satzger et al. [5] and Nafz et al. [16]) as the illustrative example.

4.3 Architecture Meta-Model

Fig. 3 shows the sco/c architecture meta-mode that supports and incorporates the bio-inspired principles mentioned before. The reason for proposing this meta-model is to point out the sco/c architecture works for systems that follow this meta-model and have its main elements.

A closer look at the meta-model shows the influence of principles 1, 2 and 3 clearly, since the meta-model is based on interacting agents with normal capabilities and general OC capabilities. Communication between agents realizes Principle 6 (i.e., micro-environment and chemicals in it).

There are two additional OC capabilities in the meta-model, named regulation and expression. The introduction of these two mandatory OC capabilities helps to realize the needed local control (Principle 7) and contribute to the self-configuration in the whole system. These two OC capabilities are defined more precisely as follows:

- The regulation capability must identify the proper o/c component configuration by activating the needed OC capabilities and deactivating the unnecessary ones. This function is similar to what happens inside each cell.
- The expression capability resolves the dependencies between OC capabilities that are identified by regulation. The expression capabilities of various agents collaborate with each other when needed.

Returning to our illustrative example, the robots are independent identical entities that can be safely considered as agents. Their capabilities are drilling, insertion and tightening. In this way, principles 1 and 2 are satisfied. The OC capabilities and activation/ deactivation of these capabilities in the example will be introduced later.

4.4 Self-Configuration for the SCO/C Architecture

The demonstration of self-configuration in the sco/c architecture requires the description of the usual behavior of the system. The scenario for sco/c can be described in short as follows. The system begins in an embryonic state in which no OC capability is active (Principle 5).

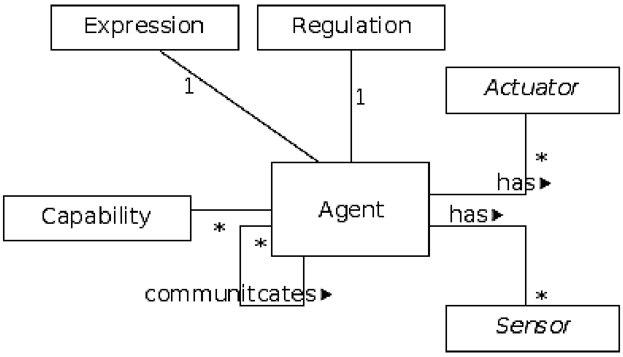


Fig. 3. The sco/c architectural meta-model

Firstly, both the regulation and expression are activated, so the local control is realized. The regulation capability identifies the OC capabilities needed to be activated. The expression capability resolves the dependencies. After that, the OC capability/capabilities that must be activated in each agent is indicated by a distributed algorithm. Finally, the desired OC capabilities are activated by the regulation capability.

Until the end of this section, the above mentioned scenario is presented with more details. The identification of the needed OC capabilities in the current sco/c architecture is in form of rules supplied by the architect in the design time (wrong rules will lead to undesired outcomes). Therefore, the validity of the mechanism is totally dependent on the mindset of the architect since the control mechanism in the sco/c architecture cannot understand the semantics of such rules. These rules have the generic form of “if-then” meaning that if a condition is matched, some capabilities are considered to be needed (i.e., an o/c architecture configuration).

The feature model (Section 2-3) is a good candidate for capturing the OC capabilities and their relationships in the form of a hierarchy. The possible configuration for the o/c component can be given through the feature diagram.

Fig. 4 shows a feature diagram for the illustrative example incorporating [5] and [16] as the two related works presented in Section 4-2 (the organic middleware [5] and the RIA controller [16]). The OCu represents the organic middleware controller. It must be mentioned that other o/c component configurations can be incorporated in the feature diagram, but we used the ones that suit our illustrative example the best. As can be seen, the o/c component mandates both observer and controller. The observer can have one of the two o/c components (invariant monitor and information pool manager). The “require” relationship indicates inter-tree relations between the OC capabilities. For example, the RIA controller can be realized using the “require” relationship between invariant monitoring and RIA controller. The final subcomponents of the RIA controller must be realized because of the mandatory relationship between the result checker and the reconfiguration algorithm.

After the identification of the OC capabilities by the regulation capability in each agent, all of the agents know that the o/c component configuration must be activated.

Next, each agent compares its OC capabilities with the needed OC capabilities for realizing the selected o/c component configuration. There might be multiple instances of each needed OC capabilities identified by the agents. In other words, many agents may have the needed OC capabilities. They are announced to the neighborhood and ultimately all the system. After that, a distributed election algorithm, such as the one introduced in [17], elects the desired OC capabilities. The algorithm denotes which OC capability in which agents must be activated. Therefore, any other OC capability except those indicated by the algorithm must be deactivated. The reason for deactivation is that there might be a previous o/c component configuration. If this deactivation does not happen, there might be another configuration active, and this might lead to unexpected results. This causes the sco/c architecture to be usable in variety of environments, i.e., if the regulation can determine the type of the o/c component configuration, it make the system operate automatically and without manual intervention. If the activation/deactivation process is completed, all the desired OC capabilities are activated, and a special configuration of the o/c architecture can be realized. After a successful configuration, the OC system starts to operate. It can be said that, in the sco/c architecture, there are two distinct self-configuration and operation stages. Self-configuration is involved with the realization of the control mechanism, while in the operation stage, the SuOC is reconfigured accordingly.

If we consider Fig. 4 as the feature diagram, the following argument can be presented for the RIA controller and the organic middleware: The former uses a centralized variant of the o/c architecture, while the latter uses a decentralized one. The decisions are centralized in the former and easier to achieve, while in the latter, the decisions are made independently and then coordinated. So, a key architectural decision would be to choose and employ a proper new configuration from these two alternative o/c architectures. As one of the configurations, we can assume the computational power of the agents in the regulation rules supplied by the architect. The computational power is chosen because the organic middleware requires that its instances run on each agent and make decisions, therefore requires higher computational power, and consequently power usage. This power usage is a major concern when it comes to general applications of pervasive or ubiquitous computing. On the other hand, the RIA controller is centralized and has more lightweight components (or in other words, less computational power) than the organic middleware.

When the decision is made and one of the variants is chosen, the required components should be identified. For instance, if we want to select the RIA controller itself, all the agents should choose the respective components: invariant monitor, reconfiguration algorithms and result checker. In our illustrative example, since the agents (robots) are identical, all of them announce the three needed OC capabilities. The distributed election algorithm will eventually specify the appropriate allocation of the OC capabilities. The expression of each robot activates the selected OC capabilities and deactivates the others. After that, the system can begin its normal operation.

5. Formal Specification and Verification

In order to present the sco/c architecture more precisely and with less ambiguity, and verify its self-configuration property formally, the sco/c architecture is specified. We also used Linear Temporal Logic (LTL) [18] for expressing invariants needed for the sco/c architecture. Our approach for verification is using model checking capabilities of the Maude formal tool [18].

5.1 Specification

Our specification is focused on the self-configuration phase because it involves all of the contributions of this paper. Specifications 1 through 5 describe the regulation and expression capabilities and the governing conditions in Maude.

Specification 1. This specification formalizes Principle 7 for the regulation capability. It is a collection of rules supplied by the architect.

$$\mathrm{Condition}\colon \mathbb{P}\mathrm{Parameter} \rightarrow \mathrm{Boolean} \qquad (1)$$

$$\mathrm{RegulationRule}\colon \mathrm{Condition} \rightarrow \mathbb{P}\mathrm{Capability} \qquad (2)$$

$$\mathrm{Regulation} == \mathbb{P}\mathrm{RegulationRule} \qquad (3)$$

$$\mathrm{regulate}\colon \mathrm{Condition} \times \mathrm{Regulation} \rightarrow \mathbb{P}\mathrm{Capability} \qquad (4)$$

Parameter represents any information (such as the number of agents or agent distribution) that can be used for decision making and selecting the OC capabilities. Condition (Declaration 1) is a function that takes a set of parameters (as a specific condition) into account and returns a Boolean value representing the validity of that condition. For example it checks if $\mathrm{ping\ time} > 10\mathrm{ms}$ and $\mathrm{bandwidth} > 64\mathrm{K}$ as two parameters constituting a specific condition holds or not. Regulation (Declaration 3) is defined as a set of RegulationRule (Declaration 2) RegulationRule defines the elements of the regulation capability in the form of a rule that specifies a proper set of capabilities for each condition. The regulate function (Declaration 4) specifies the regulation function of the regulation capability. It takes a condition and the set of RegulationRule and returns the capabilities that are needed to be activated in that condition.

Specification 2. Similarly, Expression (Declaration 5) denotes the expression capability. It shows the relationships reltype between the OC capabilities according to the feature diagram. This specification formalizes Principles 3, 4 and 7.

$$\begin{gathered}\mathrm{Expression}\colon \mathbb{P}\mathrm{Capability} \rightarrow \mathbb{P}(\mathrm{reltype} \times \mathrm{Capability}) \\ \mathrm{reltype} = \{\mathrm{mandatory}, \mathrm{optional}, \mathrm{excludes}\ \ \mathrm{or\ requires}\}\end{gathered} \qquad (5)$$

This specification formalizes the relationships discussed in Principle 4. Using reltype defined in this specification, the relationships between OC capabilities can be represented.

Apart from these specifications, additional ones are needed in order to specify operations and normal capabilities of the agents. Since we have focused on OC capabilities, the specification can be simplified by ignoring other operations and normal capabilities of the agents.

Specification 3. Based on the sco/c meta-model and principles 1, 2 and 3, the agent can now be defined (Declaration 6) regarding an instance of Regulation, an instance of Expression, a set of active OC capabilities and a set of inactive OC capabilities. The active OC capabilities represent the active/expressed OC capabilities, while the inactive OC capabilities represent the deactivated/repressed OC capabilities. These two sets of OC capabilities have no intersection. In other words, no capability can be both active and inactive at the same time; see Equation 7.

$$\mathrm{Agent}\colon \mathrm{Regulation} \times \mathrm{Expression} \times \mathbb{P}\mathrm{Capability} \times \mathbb{P}\mathrm{Capability} \qquad (6)$$

$$a = (R, E, AP, IP)\ \mathrm{where}\ a \in Agent \wedge R \in Regulation \wedge E \in Expression\ \wedge P \in \mathbb{P}Capability \wedge IP \in \mathbb{P}Capability \wedge AP \cap IP \qquad (7)$$

A few auxiliary functions are needed for simplifying the specification. They are presented in declarations 8 to 11. The disableCap and enableCap functions represent the actions of enabling and disabling capabilities, respectively. They take a set of capabilities and enable/disable them in an agent. Therefore, they return an agent with new capabilities. The filter function takes sets of pairs of relationship types (mandatory, optional, requires and excludes) and OC capabilities ($\mathrm{reltype} \times \mathbb{P}\mathrm{Capability}$) alongside a relationship type (the second argument of filter in Declaration 10) for filtering and returns the set of OC capabilities whose relationship type is the same as the type determined as the second argument of filter. For instance, this function can assist in extracting the OC capabilities that are mandatory or needed. Declaration 11 denotes a simple auxiliary function which returns all the capabilities (either active or inactive) of an agent.

$$\mathrm{disableCap}: \mathbb{P}\mathrm{Capability} \times \mathrm{Agent} \rightarrow \mathrm{Agent} \tag{8}$$

$$\mathrm{enableCap}: \mathbb{P}\mathrm{Capability} \times \mathrm{Agent} \rightarrow \mathrm{Agent} \tag{9}$$

$$\begin{aligned}\mathrm{filter}: \mathbb{P}(\mathrm{reltype} \times \mathrm{Capability}) \times \mathrm{reltype} \\ \rightarrow \mathbb{P}\mathrm{Capability}\end{aligned} \tag{10}$$

$$\mathrm{capabilitySet}: \mathrm{Agent} \rightarrow \mathbb{P}\mathrm{Capability} \tag{11}$$

Specification 4. The specification of the used election algorithm is in the form of a function (elect in Declaration 12) that returns the set of pairs of agents and the set of OC capabilities ($\mathbb{P}(\mathrm{Agent} \times \mathbb{P}\mathrm{Capability})$) denoting which OC capability or capabilities of each agent must be activated. Regarding Declaration 13, $\mathrm{isElected}$ is another auxiliary function related to the elect function that returns the elected OC capabilities ($\mathbb{P}\mathrm{Capability}$) for an agent ($\mathrm{Agent}$) through an election ($\mathbb{P}(\mathrm{Agent} \times \mathbb{P}\mathrm{Capability})$).

$$\begin{aligned}\mathrm{elect}: \mathbb{P}(\mathrm{Agent} \times \mathbb{P}\mathrm{Capability}) \times \mathbb{P}\mathrm{Capability} \\ \rightarrow \mathbb{P}(\mathrm{Agent} \times \mathbb{P}\mathrm{Capability})\end{aligned} \tag{12}$$

$$\begin{aligned}\mathrm{isElected}: \mathbb{P}(\mathrm{Agent} \times \mathbb{P}\mathrm{Capability}) \times \mathrm{Agent} \\ \rightarrow \mathbb{P}\mathrm{Capability}\end{aligned} \tag{13}$$

Having these functions in place, we are ready to define the self-configuring mechanism in form of function application. To do so, we need to declare the required variables (Declaration 14).

$$\begin{aligned}&\mathrm{c} \in \mathrm{Condition} \\ &\mathrm{agents}: \mathbb{P}\mathrm{Agent} == \{\mathrm{a}_0, \mathrm{a}_1, \dots \mathrm{a}_\mathrm{n}\}\end{aligned} \tag{14}$$

Equations 15 and 16 show a few abbreviations and variable definitions to simplify Specification 5. $\mathrm{involvedCaps}$ are the mandatory, optional and required OC capabilities involved in the sco/c architecture. Equation 17 is of particular interest. It shows the candidates of agents (and their capabilities across the system) that can take part in the sco/c architecture according to the regulation and expression (see the definition of $\mathrm{involvedCaps}$ in Definition 16). These candidates resulted as follows: for each agent, its OC capabilities (members of $\mathrm{capabilitySet}(\mathrm{a}_\mathrm{i})$) that are involved in $\mathrm{involvedCaps}$ is obtained. The excluded capabilities are used later for disabling the unnecessary OC capabilities (see Equation 19). The resulting candidates and the set of the involved capabilities are finally given to the elect function which determines the required OC capability or capabilities for activation.

$$\begin{aligned}&\mathrm{allInvolvedCaps} == \mathrm{Expression}(\mathrm{regulate}, \\ &\quad (\mathrm{c}, \mathrm{Regulation}))\end{aligned} \tag{15}$$

$$\begin{aligned}&\mathrm{involvedCaps} == \\ &\quad \mathrm{filter}(\mathrm{allInvolvedCaps}, \mathrm{mandatory}) \\ &\quad \cup \mathrm{filter}(\mathrm{allInvolvedCaps}, \mathrm{optional}) \\ &\quad \cup \mathrm{filter}(\mathrm{allInvolvedCaps}, \mathrm{required})\end{aligned} \tag{16}$$

$$\begin{aligned}&\mathrm{candidates} == \\ &\bigcup_{\mathrm{i}=0}^{\mathrm{n}} (\mathrm{a}_\mathrm{i}, \mathrm{involvedCaps} \cap \mathrm{capabiltiySet}(\mathrm{a}_\mathrm{i}))\end{aligned} \tag{17}$$

$$\mathrm{election} == \mathrm{elect}(\mathrm{candidates}, \mathrm{involvedCaps}) \tag{18}$$

Specification 5. Finally, to specify the self-configuration mechanism, the system is specified in Equation 1, and the self-configuration mechanism is shown in form of a function application. Initially, for each agent, the excluded OC capabilities are disabled, and then, the elected OC capabilities of that agent are enabled. The formed OC system has thus all the OC capabilities required for the selected o/c architecture enabled by the regulation capability of the agents.

$$\begin{aligned}&\mathrm{system} == \{\mathrm{a}_0, \mathrm{a}_1, \dots \mathrm{a}_\mathrm{n}\} \\ &\quad \mathrm{where}\ \mathrm{a}_\mathrm{i} \in \mathrm{Agent} \wedge \\ &\quad \mathrm{a}_\mathrm{i} = \mathrm{enableCap}(\mathrm{isElected}(\mathrm{election}, \mathrm{a}_\mathrm{i}), \\ &\quad \mathrm{disableCap}(\mathrm{filter}(\mathrm{allInvolvedCaps}, \\ &\quad \mathrm{excludes}), \mathrm{a}_\mathrm{i}))\end{aligned} \tag{19}$$

5.2 Verification

In order to verify the self-configuration property of the sco/c architecture, we use LTL model checking. What is important in terms of self-configuration is that the system will be eventually in a state of proper operation [4], which means that: First, an o/c component configuration has been selected. Second, the OC capabilities are successfully identified, and the agents that must activate them are specified via the election algorithm. Third, the activation/deactivation of OC capabilities is done.

Having all the above mentioned conditions, it can be said that “the system is in a valid o/c component configuration”. These phrases can be expressed in the form of an invariant (Formula 20) where system comes from Equation 19, and c comes from Declaration 14.

$$\begin{aligned}&\diamond\ \mathrm{conforms}(\mathrm{system}, \\ &\quad \mathrm{Expression}(\mathrm{regualte}(\mathrm{c}, \mathrm{Regulation}))\end{aligned} \tag{20}$$

The $\mathrm{conforms}$ function is an auxiliary function that checks the validity of the OC system. It uses a condition variable (c) that the sco/c architecture has been selected.

Therefore, the OC capabilities are extracted from the OC system and used for comparison so as to see the conformation to the valid model returned by the regulate function.

It should be noted that the validity of sco/c depends on the rules defined in the regulation. With wrong rules (such as impossible configuration or unreachable conditions), sco/c does not work, and the system cannot be configured. These conditions include applying those rules that employ non-existing OC capabilities or when the needed capabilities cannot be found in the agent and its neighborhood. Also, when none of the conditions can be evaluated to true, and sco/c cannot self-configure. The same can be said when more than one o/c architecture can be selected. In this condition, the agents will split into two or more groups each trying to achieve a specific o/c architecture. Depending on the OC capability distribution among each group, different outcomes can be expected (such as zero, one or more successful o/c architecture configuration). But, any outcome in this state cannot be accepted, and even, if it works, it will be accidental.

The verification was performed successfully for various possible scenarios using the Maude tool. The target scenarios were divided into two types. The focus of the first type was on situations in which a variant of the o/c architecture could be applied. The goal was to see whether the proper variant was selected and activated in such scenarios. The second type of scenarios included the ones in which no variant were applicable. Also, the verification was performed for impossible and unreachable configurations. In all of the scenarios, the specification of the sco/c architecture proved to be sound and correct.

6. Case Study

In this section we demonstrate the applicability of the sco/c architecture on the example illustrated in subsection 4.2 using our formal specification and verification approach. We will first specify the general form of the problem, meaning the number of tasks and robots is not limited to what has been specified in [5] and [16]. Next we will use this general form to verify the illustrative example.

6.1 General Description

Most of the specifications needed for this example have been already provided. Apart from our intention to present a case study for the application of the sco/c architecture, we also intend to specify normal operations (along with the sco/c architecture specification), resulting in a complete system specification. Specification 6 describes a generic robot in which Agent has already been introduced in Declaration 6. $\mathbb{P}\mathrm{WorkingCapability}$ is used for indicating a set of normal capabilities (sensors and actuators), and $\mathbb{P}\mathrm{Resource}$ for a set of resources that the robot is working on. This set can be $\emptyset$ when there is no resource.

Specification 6. Robot definition.

$$\mathrm{Robot} == (\mathrm{Agent} \times \mathbb{P}\mathrm{WorkingCapability} \\ \times \mathbb{P}\mathrm{Resource}) \tag{21}$$

We defined a simple behavior for each robot based on [6] and then applied the general theme discussed above. The behavior of each robot in our specification consists of three general actions, i.e., acquire, process and release (Specification 7): the resource is acquired, processed and finally released. It is important to note that no action must be done unless the sco/c architecture is formed. This guarantees the formation of the self-configuration phase.

As discussed in the previous section, the $\mathrm{conforms}$ function has the responsibility of checking the conformance between the current formed architecture and the desired one. We use this function as a guard (whose result is used as the Boolean parameter in equations 22 through 24) for all of the actions in our case study.

$\mathbb{P}\mathrm{Resource}$ indicates the set of all available resources; with each acquire operation for a resource ($\mathrm{Resource}$ as the third argument for the $\mathrm{acquire}$ function) or when a resource is completely processed, it is removed from the resource set. Because the definition of Robot has an occurrence $\mathbb{P}\mathrm{Resource}$ that indicates the resource the robot is working on, when this set changes, the Robot needs to change. Therefore, the Robot is considered as both input and output in declaration 22 to 24. By the $\mathrm{process}$ function (Declaration 23), a task is performed on a resource or resources. Finally, when the process is done or a problem happens (e.g., one of the robot tools is broken), the resource is released by the robot (Declaration 24) and added to the resource list.

Specification 7. Simple behavior for the robots.

$$\mathrm{acquire}\colon \mathrm{Boolean} \times \mathrm{Resource} \times \mathbb{P}\mathrm{Resource} \\ \rightarrow \mathrm{Robot} \times \mathbb{P}\mathrm{Resource} \tag{22}$$

$$\mathrm{process}\colon \mathrm{Boolean} \times \mathrm{Robot} \rightarrow \mathrm{Robot} \tag{23}$$

$$\mathrm{release}\colon \mathrm{Boolean} \times \mathrm{Robot} \times \mathbb{P}\mathrm{Resource} \\ \rightarrow \mathrm{Robot} \times \mathbb{P}\mathrm{Resource} \tag{24}$$

After using these behaviors and completing the required definitions (i.e., specifying $\mathrm{WorkingCapability}$, $\mathrm{Regulation}, \mathrm{Expression}$, etc.), the LTL formula 20 is verifiable. The next subsection completes these items for the illustrative example and performs the verification.

6.2 Specification and Verification of the Illustrative Example

This subsection presents definitions specific to the illustrative example.

The first step is to define the regulation which consists of two rules ($\mathrm{regulation}_1$ and $\mathrm{regulation}_2$given below). Based on the computing power of the robots (as the condition), one of the o/c variants can be selected: The OC middleware (mentioned in section 3) is more suitable for higher computational power since it needs an instance of the middleware running on each robot, making it suitable for the decentralized variant, while the RIA controller needs less computational power compared to that of the OC middleware.

$$\mathrm{computingPower}\colon \mathrm{Boolean} \tag{25}$$

$$\mathrm{regulation}_1 = (\mathrm{computingPower}, \\ \{\mathrm{Information\ Pool\ Manager}, \mathrm{Fact\ Base}, \\ \mathrm{Event\ Manager}, \mathrm{Reflex\ Manager}, \\ \mathrm{High\ Level\ Planner\ Manager}, \\ \mathrm{High\ Level\ Planner}, \mathrm{Actuator}, \\ \mathrm{Low\ Level\ Reflexes}\}) \tag{26}$$

$$\mathrm{regulation}_2 \\ = (\neg\mathrm{computingPower}, \{\mathrm{Invariant\ Monitor}, \tag{27}$$

$$\text{Reconfiguration Algorithm}, \text{Result Checker}\})$$
$$\text{Regulation} = \{\text{regualtion}_1, \text{regulation}_2\} \quad (28)$$

Definition 25 specifies computingPower as the required condition. This function returns true when the computational power is suitable for running the OC middleware; when the function returns false, it means the RIA controller should be used.

Based on the above specification, the regulate function should be called with computingPower as the first parameter. Due to space limitation, the specification of expression capability has been ignored, but it can be easily extracted from Fig. 4. Declaration 29 shows the specification of Agent used in the robot definition for the illustrative example. As can be seen, the Agent components have been replaced by definitions from this section. Resource has been defined as sequence of WorkingCapability, meaning that each task is represented by the corresponding tool. When each task is performed, the first task in the sequence is removed from Resource. An empty Resource represents a resource on which all the required tasks have been successfully performed.

$$\begin{aligned}&\text{Agent} = (\text{Regulation}, \text{Expression},\\ &\quad \text{regulate}(\text{computingPower}, \text{Regulation})),\\ &\quad \text{Regulation} - \text{regulate}(\text{computingPower},\\ &\quad \text{Regulation}))\end{aligned} \quad (29)$$
$$\text{WorkingCapability} = \{\text{drill}, \text{insert}, \text{tighten}\} \quad (30)$$
$$\text{Resource} = \text{seq}(\text{WorkingCapability}) \quad (31)$$

Now robots for the illustrative example can be defined using Declaration 32.

$$\text{r}_1, \text{r}_2, \text{r}_3\text{: Agent} \quad (32)$$

As for the verification, the LTL formula 20 was verified using the Maude tool. Also, verification was performed after the self-configuration phase in order to verify the conforms function as the guard of declarations 22 through 24. The verification phase showed that when the condition of the o/c component configuration changed, the system stopped operation, configuration was chosen, and then the system resumed normal operation. In cases that were designed for impossible operation, as expected, the system stopped all operations.

7. Conclusions and Future Work

In this paper, the sco/c architecture which uses the idea of cell differentiation has been presented in order to achieve self-configuration in the o/c component level. In order to support this idea, an architectural meta-model that considers the OC system as a collection of agents with some capabilities has been proposed. Among these capabilities, there are OC capabilities representing the capabilities that can perform operations related to the o/c architecture. Also, these capabilities are responsible for the self-configuration in the level of architecture itself. The sco/c architecture uses some rules provided by the architect based on the parameters of the system or environment. This architecture is then configured, and finally, the system operates. However, we believe that the rules should be adapted accordingly by considering the prior executions. In biology this notion is called genetic memory [9]. As a future work, we are planning to use the mechanisms related to this notion in order to improve the bio-inspiration and to step closer to living systems. Also, we consider the use of a simple ontology in the OC systems so as to create a semantic base. This can help to create a knowledge-based self-awareness that can assist greatly in cases like the selection of the proper o/c component configuration in the sco/c architecture. This potentially can increase the interoperability between organic systems. This can be especially useful when two or more ubiquitous or pervasive systems are needed to cooperate.

References

[1] J. O. Kephart and D. M. Chess, “The vision of autonomic computing,” Computer, vol. 36, no. 1, pp. 41–50, 2003.

[2] C. Müller-Schloer, H. Schmeck, and T. Ungerer, Eds., Organic Computing - A Paradigm Shift for Complex Systems. Springer, 2011.

[3] U. Richter, M. Mnif, J. Branke, C. Müller-Schloer, and H. Schmeck, “Towards a generic observer/controller architecture for Organic Computing.,” GI Jahrestag. 1, vol. 93, pp. 112–119, 2006.

[4] A. Berns and S. Ghosh, “Dissecting self-* properties,” in Self-Adaptive and Self-Organizing Systems, 2009. SASO’09. Third IEEE International Conference on, 2009, pp. 10–19.

[5] B. Satzger, A. Pietzowski, W. Trumler, and T. Ungerer, “Using automated planning for trusted self-organising organic computing systems,” in Autonomic and Trusted Computing, Springer, 2008, pp. 60–72.

[6] H. Seebach, F. Nafz, J.-P. Steghöfer, and W. Reif, “How to Design and Implement Self-organising Resource-Flow Systems,” in Organic Computing—A Paradigm Shift for Complex Systems, Springer, 2011, pp. 145–161.

[7] M. Salehie and L. Tahvildari, “Self-adaptive software: Landscape and research challenges.,” TAAS, vol. 4, no. 2, Jul. 2009.

[8] H. Schmeck, C. Müller-Schloer, E. Çakar, M. Mnif, and U. Richter, “Adaptivity and self-organization in organic computing systems,” ACM Trans Auton Adapt Syst, vol. 5, no. 3, pp. 10:1–10:32, Sep. 2010.

[9] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, Molecular Biology of the Cell, 4th ed. Garland, 2002.

[10] K. Kang, S. Cohen, J. Hess, W. Novak, and S. Peterson, “Feature-Oriented Domain Analysis (FODA) Feasibility Study,” Software Engineering Institute, Carnegie Mellon University, CMU/SEI-90-TR-21, 1990.

[11] M. Riebisch, “Towards a more precise definition of feature models,” Model. Var. Object-Oriented Prod. Lines, pp. 64–76, 2003.

[12] K. Lee and K. C. Kang, “Feature dependency analysis for product line component design,” in Software Reuse: Methods, Techniques, and Tools, Springer, 2004, pp. 69–85.

[13] U. Brinkschulte, M. Pacher, and A. Von Renteln, “Towards an artificial hormone system for self-organizing real-time task allocation,” in Software Technologies for Embedded and Ubiquitous Systems, Springer, 2007, pp. 339–347.

[14] M. Roth, J. Schmitt, R. Kiefhaber, F. Kluge, and T. Ungerer, “Organic Computing Middleware for Ubiquitous Environments.,” in Organic Computing, C. Müller-Schloer, H. Schmeck, and T. Ungerer, Eds. Springer, 2011, pp. 339–351.

[15] W. Brockmann, E. Maehle, K.-E. Grosspietsch, N. Rosemann, and B. Jakimovski, “ORCA: An organic robot control architecture,” in Organic Computing—A Paradigm Shift for Complex Systems, Springer, 2011, pp. 385–398.

[16] F. Nafz, J.-P. Steghöfer, H. Seebach, and W. Reif, “Formal modeling and verification of self-* systems based on observer/controller-architectures,” in Assurances for Self-Adaptive Systems, Springer, 2013, pp. 80–111.

[17] V. C. Barbosa, An introduction to distributed algorithms. MIT Press, 1996.

[18] M. Clavel, F. Dur’an, S. Eker, P. Lincoln, N. M. Oliet, J. Meseguer, and C. Talcott, All About Maude - A High-Performance Logical Framework: How to Specify, Program, and Verify Systems in Rewriting Logic. Springer, 2007.

Ali Tarihi received his BS and MS degrees in Software Engineering from Shahid Beheshti University, Tehran, Iran. He is currently a Ph.D student at the Computer Science and Engineering Faculty, Shahid Beheshti University, Tehran, Iran.

Hassan Haghighi received his BS, MS and PhD degrees in Software Engineering from Sharif University of Technology, Tehran, Iran. He is currently an assistant professor at the Computer Science and Engineering Faculty, Shahid Beheshti University, Tehran, Iran.

Fereidoon Shams Aliee received his BS and MS degrees from Shahid Beheshti University and Sharif University of Technology, respectively. He received his Ph.D in Software Engineering from Department of Computer Science, Manchester University, UK. He is currently an associate professor at the Computer Science and Engineering Faculty, Shahid Beheshti University, Tehran, Iran.

Safe Use of the Internet of Things for Privacy Enhancing

Hodjat Hamidi*
Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
h_hamidi@kntu.ac.ir

Received: 28/Feb/2016 Revised: 27/May/2016 Accepted: 18/Jul/2016

Abstract

New technologies and their uses have always had complex economic, social, cultural, and legal implications, with accompanying concerns about negative consequences. So it will probably be with the IoT and their use of data and attendant location privacy concerns. It must be recognized that management and control of information privacy may not be sufficient according to traditional user and public preferences. Society may need to balance the benefits of increased capabilities and efficiencies of the IoT against a possibly inevitably increased visibility into everyday business processes and personal activities. Much as people have come to accept increased sharing of personal information on the Web in exchange for better shopping experiences and other advantages, they may be willing to accept increased prevalence and reduced privacy of information. Because information is a large component of IoT information, and concerns about its privacy are critical to widespread adoption and confidence, privacy issues must be effectively addressed. The purpose of this paper is which looks at five phases of information flow, involving sensing, identification, storage, processing, and sharing of this information in technical, social, and legal contexts, in the IoT and three areas of privacy controls that may be considered to manage those flows, will be helpful to practitioners and researchers when evaluating the issues involved as the technology advances.

Keywords: Security Issues; IoT; Information; Technology Advances; Privacy Enhancing.

1. Introduction

Security issues are central in Internet of Things as they may occur at various levels, investing technology as well as ethical and privacy issues. To ensure security of data, services and entire IoT system, a series of properties, such as confidentiality, integrity, authentication, authorization, non-repudiation, availability, and privacy, must be guaranteed [1]. This is extremely challenging due to the Internet of Things environmental characteristics. In the past privacy was of relatively little concern because location information was not pervasively and continuously available. Now that technology has radically altered information availability, privacy of location is closely tied to controlling access to this information, and people want to be in control of the information availability [2-3]. Privacy preferences are now quite well studied in the context of users carrying mobile devices [4] but not extended through an IoT context where device-to-device communication can carry location information far beyond users' awareness. Privacy concerns are becoming an increasingly critical issue in the IoT [5]. Without assurance of privacy in a world of interconnected sensors and systems, users will be unwilling to adopt these new technologies [6]. The International Telecommunications Union report on the Internet of Things notes that "Concerns about privacy and data protection are widespread, particularly as sensors and smart tags can track a user's movements, habits, and preferences on a perpetual basis." [7] Despite its relevance and importance, privacy is not yet receiving adequate attention in the enthusiasm to exploit the technical capabilities of the IoT. A recent survey of IoT literature covering 127 journal and conference papers [8] finds only nine security and three privacy-related documents in its category of IoT challenges [9] [5] [10]. A recent survey of IoT context aware computing describes security and privacy as a major concern, yet finds only 11 of 50 surveyed research prototypes incorporating security and privacy functionality [11].

The paper is organized as follows: Section 2 presents the components in the Internet of Things. In Section 3, the security in IoT and data confidentiality is explained. In Section 4, theory concepts of privacy in IoT is introduced. Phases and associated privacy for IoT is discussed in section 5. The challenges in IoT: privacy and security are presented in Section 6. The privacy and humanness is discussed in section 7. Section 8 gives the discussion of the study. Conclusion is given in Section 9.

2. Components in the Internet of Things

The IoT vision enhances connectivity from "any-time, any-place" for "any-one" into "any-time, any-place" for "any-*thing*" [12]. Once these things are plugged into the network, more and more smart processes and services are possible which can support our economies, environment, security and health.

Fig.1 provides a view of the IoT ecosystem [13]. Things could be tagged, and through scanners, identified, and the relevant location information could be

* Corresponding Author

communicated. Similarly, networked things with sensors become smaller, weaving themselves into our daily lives, while sensor and actuator networks act on the local environment, communicating status and events to a higher level service. Smart things sense activity and status, linking it to the IoT. Middleware and frameworks enabling application and service development which utilise data as received from (or about) things, most often living in the cloud provide the capability to add intelligence resulting in better services, which ultimately impact on the environment.

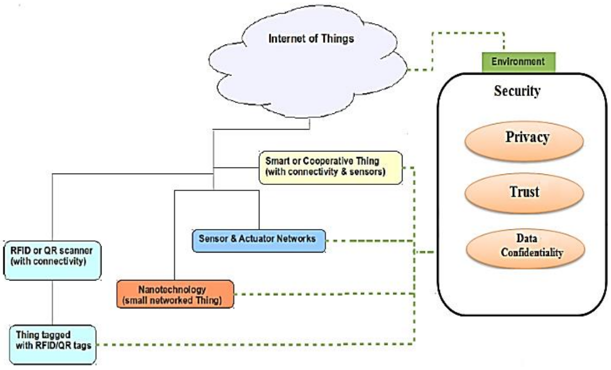

Internet of Things
Environment
Security
Privacy
Trust
Data Confidentiality
Smart or Cooperative Thing (with connectivity & sensors)
Sensor & Actuator Networks
RFID or QR scanner (with connectivity)
Nanotechnology (small networked Thing)
Thing tagged with RFID/QR tags

Fig. 1. Components in the Internet of Things

3. Security in IoT and Data Confidentiality

Security represents a critical component for enabling the widespread adoption of IoT technologies and applications. Security is divided into three parts (Fig. 2): (1) Data confidentiality, (2) privacy and (3) trust

3.1 Data Confidentiality

Data confidentiality represents a fundamental issue in IoT scenarios, indicating the guarantee that only authorized entities can access and modify data. The main research challenges for ensuring data confidentiality in an IoT scenario relate to: (1) Definition of suitable mechanisms for controlling access to data streams generated by IoT devices. (2) Definition of an appropriate query language for enabling applications to retrieve the desired information out of a data stream. (3) Definition of a suitable smart objects' identity management system.

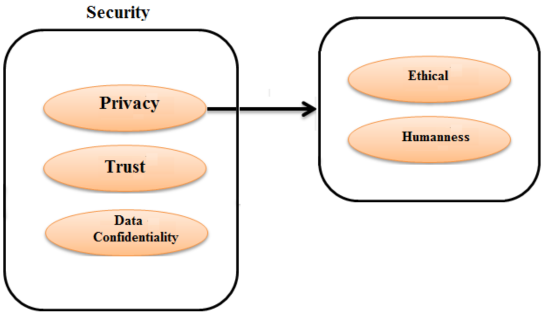


Fig. 2. Security is divided into three parts

4. Theory Concepts of Privacy in IoT

This paper follows and borrows from prior work investigating issues in the development of privacy theory general [14], and extends it to the particular environment of the IoT. Theory concepts has five desirable goals [15]:

1. A method of organizing and categorizing "things," a typology;
2. Predictions of future events;
3. Explanations of past events;
4. A sense of understanding about what causes events; and occasionally mentioned as well:
5. The potential for control of events.

At this early stage we can hardly purport to fully explain and predict the eventual evolution of the recently-emerged IoT, let alone control it—however we can begin to organize and categorize important "things" such as components and concepts.

Consistent with the above, a number of strategies may be used to construct theories, one of which is a classificatory strategy seeking a taxonomy of elements both within and outside the phenomenon [16]. In early stages of theory construction, classification strategies are particularly important and a prerequisite to other strategies [17]. This method follows recommendations from related fields [18], emphasizing discovery and description, where key research questions are "Is there something interesting enough to justify research?" and "What are the key issues?" in both cases with categorization suggested as a procedure to be used [19]. The methods described below will attempt to discover, classify, and describe a number of key issues that relate the IoT and big data to location privacy, and justify the need for additional research.

4.1 Privacy

The privacy may be viewed from many conceptual perspectives [20] and in the context of the present work related to the IoT and big data, we will consider it from an informational privacy perspective.

The informational perspective is key to most privacy theories in a technological context, describing privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others." [21]. In keeping with this approach, we will look at location privacy in terms of information flows, from sensing to use and including a number of other activities typically in between (including more complex interactions between flows).

Tables 1 and 2 use the five phases of information flow enumerated in Table 1 and identifies example privacy controls for each phase. The five phases extend early work from more than 45 years ago identifying three phases of input, storage and output [22]. They also extend five phases discussed in [23] by explicitly adding the "processing" phase to acknowledge the important of inference capabilities and data analytical techniques that may deduce location from other available evidence.

The privacy-enhancing controls fall into technical, social, and legal measures, represented in columns of the table. Technical controls are those that control the actual processing of the information and may block, filter, modify, etc. that information. Examples include authenticating, blocking, encrypting, and other privacy protections for RFID tags [24]. Social controls affect privacy information through the influence of accepted business practices, social norms, and similar nontechnical means. These include not only such things as formal privacy policies from system providers, but also behaviors of system users, which have been found to vary considerably according to context such as who the information is exchanged with, whether the person is at home or in a public place, and what means is used to share the information [25]. Legal measures are those that impose formal prohibitions or regulations on activities related to location information flows. These vary greatly by region. In the EU the privacy Directive directly addresses location privacy, while in the US federal law addresses location only indirectly and incompletely [26].

5. Phases and Associated Privacy for IoT

The phases and associated privacy controls are described:

(1) Sensing may be technically blocked by any means that prevents signal transmission or reception.

This includes RFID-blocking wallets, RF blocker tags generating simulated or false RFID tags, etc.

(2) Identification in the IoT has already received significant attention [27] [28]. Legal enforcement of anonymity is almost universally expected and enforced in particular contexts such as election ballot casting.

(3) Storage privacy is enabled through several technical methods, including merely not providing a storage facility and encryption of any stored data. The Snap chat service was touted as an ephemeral means of photo sharing, but was quickly and easily defeated [29]. Social control of stored information is often accomplished (with varying degrees of success and user satisfaction) through user privacy settings in social media. Various jurisdictions may enforce formal legal restrictions on the type, amount, and duration of stored data. A "right to quantitative privacy" has even been proposed [30].

(4) Processing phase technical privacy includes a number of design principles that also apply to other phases [31] and various anonymizing and privacy-enhancing and privacy-preserving technologies [32-34]. It may also be affected on a social and free market level in software terms of service agreements. Formal legal measures include prohibitions or restrictions on database matching and sharing of information between commercial entities.

(5) Sharing phase privacy may be technically implemented by restricting the communications channels available, e.g., not implementing or turning off facilities such as Bluetooth and Wi-Fi. Social measures are largely the responsibility of users to control application settings and follow recommended norms for appropriate sharing. Legal controls for sharing have recently received significant attention—for example the US Federal Trade Commission has just recommended that Congress give consumers more control over the data brokerage industry [35-37] and European courts have required that search engines implement a "right to be forgotten" [38].

Table 1. Information flow phases and associated privacy controls for IoT

Phases	Methods
Identification	-Unique identifier detection -Facial recognition -Vehicle license plate recognition
Processing	-Self-contained inference -Communication and matching
Sharing	-Intentional -Unintentional
Storage	-Object data -Meta data
Sensing	-Triangulation -Scene analysis -Proximity -Indirect inference

Table 2. Privacy measures for IoT

Phases	Social privacy	Technical privacy	Legal privacy
Identification	Anonymous letters to newspaper editors or postings to online discussion forums	address randomization	"Secret" ballots for voting
Processing	Vendor-customer terms of service	Privacy-enhancing technologies: anonymizing, etc.	Restrictions of database matching
Sharing	User and application sharing settings	Restriction or no provision of communication facilities	The "right to forget" Data broker restrictions
Storage	User social media privacy settings	No physical storage -Encryption -Ephemeral storage	Formal limits on amount and duration of stored data
Sensing	Socially acceptable uses for Google Glass [39]	RF blocking wallets	Prohibition of cell phone and camera use at customs [40]

6. Challenges in IoT: Privacy and Security

This section discusses challenges in IoT development by enterprises. As with any disruptive innovation, the IoT will present multiple challenges to adopting enterprises. For example, due to the explosion of data generated by IoT machines, in [40] suggested that data centers will face challenges in security and consumer privacy. This section discusses two technical and managerial challenges: privacy and security.

6.1 Privacy Challenge

As is the case with smart health equipment and smart car emergency services, IoT devices can provide a vast amount of data on IoT users' location and movements, health conditions, and purchasing preferences-all of which can spark significant privacy concerns. Protecting privacy is often counter-productive to service providers in this scenario, as data generated by the IoT is key to improving the quality of people's lives and decreasing service providers' costs by streamlining operations. The IoT is likely to improve the quality of people's lives. According to the 2014 TRUST Internet of Things Privacy Index, only 22% of Internet users agreed that the benefits of smart devices outweighed any privacy concerns. While the IoT continues to gain momentum through smart home systems and wearable devices, confidence in and acceptance of the IoT will depend on the protection of users' privacy.

6.2 Security Challenge

As a growing number and variety of connected devices are introduced into IoT networks, the potential security threat escalates. Although the IoT improves the productivity of companies and enhances the quality of people's lives, the IoT will also increase the potential attack surfaces for hackers and other cyber criminals. IoT devices have vulnerabilities due to lack of transport encryption, insecure Web interfaces, inadequate software protection, and insufficient authorization. On average, each device contained 25 holes, or risks of compromising the home network. Devices on the IoT typically do not use data encryption techniques.

Some IoT applications support sensitive infrastructures and strategic services such as the smart grid and facility protection. Other IoT applications will increasingly generate enormous amounts of personal data about household, health, and financial status that enterprises will be able to leverage for their businesses. Lack of security and privacy will create resistance to adoption of the IoT by firms and individuals. Security challenges may be resolved by training developers to incorporate security solutions (e.g., intrusion prevention systems, firewalls) into products and encouraging users to utilize IoT security features that are built into their devices.

The evolution of IoT technologies (e.g., chips, sensors, wireless technologies) is in a hyper accelerated innovation cycle that is much faster than the typical consumer product innovation cycle. There are still competing standards, insufficient security, privacy issues, complex communications, and proliferating numbers of poorly tested devices. If not designed carefully, multi-purpose devices and collaborative applications can turn our lives into chaos. To prevent chaos in the hyper-connected IoT world, businesses need to make every effort to reduce the complexity of connected systems, enhance the security and standardization of applications, and guarantee the safety and privacy of users anytime, anywhere, on any device.

Beyond the security challenges mentioned in other parts of this document, we identify specific additional challenges here:

6.2.1 Diverse, Interacting, Potentially Unsecure Devices:

IoT raises a wide range of serious security challenges, since many IoT devices interact closely with the physical world. Recent news has highlighted many opportunities for attack on networked cars, power stations, and implanted medical devices. The security problem is exacerbated by the fact that many IoT devices may be built by companies that have little expertise in security, using potentially old operating systems and libraries that are not fully patched. Furthermore, if a device relies on open-source software with vulnerabilities, updating the firmware on such devices can be difficult.

6.2.2 Devices that Misrepresent Themselves:

Another risk lies in the potential for these diverse devices to be intentionally programmed to "cheat" as was the recent case where Volkswagen was found to have programmed their software to cheat on emissions tests [16]. By cheating, we mean any action that intentionally misrepresents the product's behavior for the purpose of deceiving regulators or consumers. Examples of such cheating might be misrepresenting network bandwidth usage or performance on benchmarks. As we cede more control to these devices the need to regulate them will increase, which will give manufacturers more temptation to cheat. Technologies, procedures, and policies are needed to allow inexpensive and effective auditing of the software in such devices, including methods to specify the expected correct behavior and solutions that allow for inspection of the product source code.

6.2.3 Security Threats from Ubiquitous Devices:

In a world where we are surrounded by IoT devices, the ability to limit our exposure to them decreases. If a desktop computer becomes infected, we can reboot it, run an anti-virus program, and hope the problem goes away. If one or more devices in a network of IoT devices is compromised, it may be both very difficult to know what device has been compromised or how to fix the problem to restore the overall system security. Consider how current ransom ware, which holds our data hostage, might be transformed to an attack that requires us to pay money to enter our own house or turn on the heat. Research on systematic methods for restoring IoT systems from a known good state is needed as well as tools to isolate and correct individual compromised components within the distributed system.

6.2.4 System-wide Security Abstractions:

Programming languages have evolved to incorporate features that increase productivity and reduce classes of errors. For example, Java and C# have features that prevent errors such as buffer overflows by construction – all valid programs are correct with respect to memory safety. Next-generation IoT systems, that involve physical interaction, need to have a new generation of system-wide properties (e.g., to guarantee physical safety) that are correct by construction and checked automatically. These properties involve major improvements in our ability to reason about the interaction between the software in the system and the physics of their real-world actions.

7. Privacy and Humanness

7.1 Ethical Challenges

In Internet of Things exchange environments, there are more data that can be used to define and to influence people. Will these data, which in digital form are coded as strings of zeroes and ones, lead marketers to view consumers strictly as data, slotting them into fixed categories and treating them with sterile precision in accordance to their assignment? And through the acquisition and sharing of these data-perhaps in the end, without much choice by those from whom the data are sourced-will consumers relinquish important aspects that define their humanness, and thus feel less satisfaction? In the context of mortality and being human, Gawande [40] draws upon Dworkin (1986), [41], and his perspective on autonomy to put forth, ‘‘we want to retain the autonomy-the freedom-to be the authors of our lives. This is the very marrow of being human. All we ask is to be allowed to remain the writers of our own story. That story is ever changing. We want to retain the freedom to shape our lives in ways consistent with our character and loyalties.’’ Although he was writing about mortality when framing that ‘‘the battle of being mortal is the battle to maintain the integrity of one’s life’’ [40], we believe it applies when one, through technology and associated data, can be increasingly represented, influenced, and con-trolled and as a result have choices censored. The autonomy of one’s data and thus one’s self should be respected and the individual should be provided freedom of control. We believe that the human condition calls for and requires sufficient privacy. Indeed, privacy and all that it represents or entails is a sine qua non of humanness. Without it, consumers may feel like- and become-empty souls and vessels through which organizations derive profits.

7.2 The Human Condition

Thinking more like a technologist does not suggest thinking less like or about people. In fact, we argue that it becomes more important to consider the human condition when designing technology-based solutions. We believe there is a tendency to use technology as a mass market solution to a problem in which solutions address the major issues or are believed to be robust enough to provide a reasonable solution to any problem. However, such solutions may overlook smaller details or individual preferences that may comprise the long tail and therefore may be less satisfactory than imagined. For example, consider automated phone call-in systems where customers ‘‘Listen closely as options may have changed . . . Press 1 for . . . Press 2 for . . . .’’ It seems like a grand way to handle a large volume of calls on a variety of topics. However, in application, too many consumers may be frustrated by such systems. Brands take a hit when this happens. A method intended to enhance customer service can ironically result in annoying customers. Thinking more fully through the human condition will yield more effective solutions.

8. Discussion

The nature of IoT means that researchers can now “lurk” in wait for what are, in essence, ready-made data sets [19]. However, the speed with which these new sources of data have emerged, as well as the increasingly imaginative ways that researchers are using them, risked running ahead of the development of an appropriate ethical framework for their use.

Ethicists have recognised that they face a challenge in determining how to transfer traditional deontological principles into the world of IoT, addressing the duties and obligations of the researcher, as well as how to deal with concepts such as utilitarianism, feminism, and communitarianism [20]. As research on material published on the internet involves no direct contact between the subject and the researcher. It avoids one of the problems facing much qualitative research, namely that of interviewer bias, whereby what is said is influenced by the researcher. However, the absence of such contact creates other problems, in particular those relating to informed consent and protection of the subject. This commentary considers two of the major issues in the ethics of IoT research; the difference between public and private space and the right to anonymity.

Now that people routinely share detailed information on all aspects of their lives, including embarrassing anecdotes and even incriminating photographs on social media, there are questions as to what online privacy actually means. One approach is to apply the ethic of reciprocity, or Golden Rule, whereby the researcher asks how they would feel if the roles were reversed [21].

The challenge then is to operationalize this principle. How do people’s expectations of privacy change depending on the type of IoT they are using and what are the consequences for researchers’ ethical obligations [28]?

This discussion recognises that the concept of privacy is inherently complicated and there is a need to understand how individuals will respond to violations in different contexts [29].

A related issue is that of anonymity. Anonymity is a fundamental right of subjects of research. It underpins the potentially fragile trust between the subject and the researcher and is integral to consent and provision of information as well as being a manifestation of the respect in which the researcher holds the subject in front of the computer screen. This creates many additional ethical considerations for the researcher [28,31].

9. Conclusion

The Internet of Things is the connection – via the internet – of objects from the physical world that are equipped with sensors, actuators and communication technology. IoT systems will create dramatic business opportunities and provide great benefits to individuals and society. For these systems to succeed, they must be secure, robust, and usable by humans. Progress has been made on improving the security of existing systems but IoT systems require even higher quality and introduce new complexities.

Privacy and security are the important aspect for Internet of Things (IoT) deployments. In this paper, we provided an overview of the privacy of IoT technologies, and a number of research challenges has been identified, which are expected to become major research trends in the next years. Several significant obstacles remain to fulfill the IoT vision, among them privacy. Indeed, realizing the IoT vision is likely to spark novel and ingenious malicious models. The challenge is to prevent the growth of such models or at least to mitigate and limit their impact. Meeting this challenge requires understanding the characteristics of things and the technologies that empower the Internet of Things. When every object in our daily life is connected to the Internet, they must be secure. Although the IoT improves the productivity of companies and enhances the quality of people's lives, the IoT will also increase the potential attack surfaces for hackers and other cyber criminals. Lack privacy will create resistance to adoption of the IoT by firms and individuals.

For future research, the following questions can be considered: (1) what are the types and levels of behaviors adopted by users as a result of their privacy concerns, and why are these adopted? (2) What are the differences in awareness, concerns, and behaviors of the general public versus business entities related to privacy?

Moreover, Support research that addresses the core underlying scientific and engineering principles dealing with large-scale issues, networking, security, privacy, real-time, and the other key questions raised in this paper.

References

[1] E. Rekleitis, P. Rizomiliotis, and S. Gritzalis, "A Holistic Approach to RFID Security and Privacy," Proc. 1st Int'l Workshop Security of the Internet of Things (SecIoT 10), Network Information and Computer Security Laboratory, 2010; www.nics.uma.es/seciot10/files/pdf/rekleitis_seciot10_paper.pdf.

[2] D.Clark, 'Internet of Things' in reach: Companies rush into devices like smart door locks, appliances, but limitations exist. The Wall Street Journal. Retrieved April 3, 2015, from http://www.wsj.com/articles/SB100014240527 02303640604579296580892973264

[3] L. Ding, P. Shi, and B. Liu, "The clustering of Internet, Internet of things and social network," in Proc. 3rd Int. Symp. KAM, Wuhan, China, 2010.

[4] M. Nitti, R. Girau, L. Atzori, A. Iera, and G. Morabito, "A subjective model for trustworthiness evaluation in the social Internet of things," in Proc. IEEE 23rd Int. Symp. PIMRC, Sydney, NSW, Australia, 2012, pp. 18–23.

[5] F. Bao and I.-R. Chen, "Trust management for the Internet of things and its application to service composition," in Proc. IEEE Int. Symp. Wow Mom, San Francisco, CA, USA, 2012, pp. 1–6.

[6] C. Occhiuzzi, C. Vallese, S. Amendola, S. Manzari, and G. Marrocco, "NIGHT-Care: A passive RFID system for remote monitoring and control of overnight living environment," Procedia Computer Science, vol. 32, 2014, pp. 190 – 197.

[7] B.Xu, L. D. Xu, H. Cai, C. Xie, J. Hu, and F.Bu, "Ubiquitous Data Accessing Method in IoT-Based Information System for Emergency Medical Services, " IEEE Transaction on Industrial Informatics, Vol. 10, No. 2, May 2014.

[8] A. Sarma and J. Girão, "Identities in the Future Internet of Things," Wireless Personal Comm., Mar. 2009, pp. 353-363.

[9] J. Sen, "Privacy Preservation Technologies in Internet of Things," Proc. Int'l Conf. Emerging Trends in Mathematics, Technology, and Management, 2011; http://arxiv.org/ftp/arxiv/papers/1012/1012.2177.pdf.

[10] G. Broenink, "The Privacy Coach: Supporting Customer Privacy in the Internet of Things," Proc. Workshop on What Can the Internet of Things Do for the Citizen? (CIOT 2010); Radboud University, May 2010; http://dare.ubn.ru.nl/bitstream/2066/83839/1/83839.pdf.

[11] S. Radomirovic, "Towards a Model for Security and Privacy in the Internet of Things," Proc. 1st Int'l Workshop on the Security of the Internet of Things (SecIoT 10), Network Information and Computer Security Laboratory, 2010; www.nics.uma.es/seciot10/files/pdf/radomirovic_seciot10_paper.pdf.

[12] B. D. Weinberg, G. R. Milne, Y. G. Andonova, F. M. Hajjat, "Internet of Things: Convenience vs. privacy and secrecy," Business Horizons," Vol.58, No. 6, November-December, 2015, pp.615-624.

[13] M.Henze, L.Hermerschmidt, D. Kerpen, R.Häußling, B. Rumpe, K. Wehrle, "A comprehensive approach to privacy in the cloud-based Internet of Things," Future Generation Computer Systems, Vol.56, March 2016, pp. 701-718.

[14] R. H. Weber, "Internet of things: Privacy issues revisited," Computer Law & Security Review, Vol. 31, No. 5, October 2015, pp. 618-627.

[15] A.Botta, W. de Donato, V. Persico, A.Pescapé, "Integration of Cloud computing and Internet of Things: A survey Future Generation Computer Systems," Vol. 56, March 2016, pp.684-700.

[16] A. Antonić, M. Marjanović, K. Pripužić, I.P. Žarko, "A mobile crowd sensing ecosystem enabled by CUPUS: Cloud-based publish/subscribe middleware for the Internet of Things," Future Generation Computer Systems, Vol. 56, March 2016, pp.607-622.

[17] R.Neisse, G.Steri, I. N. Fovino, G. B.SecKit, "A Model-based Security Toolkit for the Internet of Things Computers & Security," Vol. 54, October 2015, pp. 60-76.

[18] I. Lee, K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," Business Horizons, Vol.58, No. 4, July–August 2015, pp. 431-440.

[19] Texas Instruments. (2014). "Application areas for the Internet of Things," Retrieved April 3, 2015, from http://www.ti.com/

[20] A.Cavoukian, J.Jonas, "Privacy by design in the age of big data. Information and Privacy Commissioner of Ontario," Canada. 2012. Available at https://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf

[21] G. R. Milne, "Digital privacy in the marketplace," New York: Business Expert Press. 2015.

[22] I.Rubinstein, “Regulating privacy by design,” Berkeley Technology Law Journal, Vol. 26, No.3, 2011, pp.1409-1456.

[23] H. Packard, “HP study reveals 70 percent of Internet of Things devices vulnerable to attack,” July 29 2014. Retrieved from http://www8.hp.com/us/en/hp-news/press-release. Html? Id=1744676#.VOTykPnF-ok

[24] L. Atzori, A. Iera, and G. Morabito, “The Internet of things: A survey,” Comput. Netw, vol. 54, no. 15, pp. 2787–2805, 2010.

[25] P. Mendes, “Social-driven Internet of connected objects,” in Proc. Interconn. Smart Objects with the Internet Workshop, Lisbon, Portugal, 2011.

[26] Z. Yan, P. Zhang, A. V. Vasilakos, “A survey on trust management for Internet of Things,” Journal of Network and Computer Applications, Volume 42, June 2014, pp. 120-134.

[27] Y. Challal, E. Natalizio, S. Sen, A. M. Vegni, “Internet of Things security and privacy: Design methods and optimization, Ad Hoc Networks,” Vol.32, September 2015, pp.1-2.

[28] J. Lee, S.Oh, J. W. Jang, “A Work in Progress: Context based Encryption Scheme for Internet of Things,” Procedia Computer Science, Vol. 56, 2015, pp. 271-275.

[29] V.H. Jeroen. “Fact sheet-Ethics Subgroup IoT -Version 4.0. Conclusions of the Internet of Things public consultation,” 2013. Available at https://ec.europa.eu/digital-agenda/en/news/conclusions-internet-things-public-consultation

[30] S. H. Rebecca. “Europe’s policy options for a dynamic and trustworthy development of the Internet of Things,” SMART 2012/0053, RAND Corp., European Union 2013, 2013.

[31] A. Samani, H. H. Ghenniwa, A.Wahaishi, “Privacy in Internet of Things: A Model and Protection Framework,” Procedia Computer Science, Vol. 52, 2015, pp. 606-613.

[32] Value Ageing’ project. “Incorporating European Fundamental Values In to ICT for Ageing: A vital political, ethical, technological, and industrial challenge,” Ref. online: www.valueageing.eu

[33] R. H. Weber, “Internet of Things–New security and privacy challenges”. Computer Law & Security Review, Vol. 26, No.1, 2010, pp. 23-30.

[34] C. M. Medaglia, A.Serbanati, “An overview of privacy and security issues in the internet of things,” In The Internet of Things, Springer New York, 2010, pp. 389-395..

[35] J. S.Kumar, D. R. Patel, “A survey on Internet of Things: security and privacy issues,” International Journal of Computer Applications, 2014, Vol. 90, No.11.

[36] S. Sicari, A. Rizzardi, L.A. Grieco, A. Coen-Porisini, “Security, privacy and trust in Internet of Things: The road ahead,” Computer Networks, Vol.76, 2015, pp.146-164.

[37] Q, Jing, A.V. Vasilakos, J.Wan, J. Lu, D, Qiu, “Security of the internet of things: Perspectives and challenges,” Wireless Networks, Vol. 20, No.8, 2014, pp. 2481-2501.

[38] H.Suo, J.Wan, C. Zou, J. Liu, “Security in the internet of things: a review,” In Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on 2012, pp. 648-651.

[39] S. Chabridon, R. Laborde, T. Desprats, A. Oglaza, P.Marie, S.M. Marquez, “A survey on addressing privacy together with quality of context for context management in the internet of things,” Annals of telecommunications-annales des télécommunications, Vol.69, No.1, 2014, pp.47-62.

[40] R.Dworkin, “Autonomy and the demented self,” The Milbank Quarterly, Vol.64, No.2, pp.4-16.

[41] A.Gawande, “Being mortal: Medicine and what matters in the end,” New York: Metropolitan Books, Henry Holt & Company, 2014.

Hodjatollah Hamidi born 1978, in shazand Arak, Iran, He got his Ph.D in Computer Engineering. His main research interest areas are Information Technology, Fault-Tolerant systems (fault-tolerant computing, error control in digital designs) and applications and reliable and secure distributed systems and E- Commerce. Since 2013 he has been a faculty member at the IT group of K. N. Toosi University of Technology, Tehran Iran. Information Technology Engineering Group, Department of Industrial Engineering, K. N. Toosi University of Technology.

Efficient Land-cover Segmentation Using Meta Fusion

Hadi Mahdipour Hossein Abad*
Department of Marine Engineering, Khorramshahr University of Marine Science and Technology, Khorramshahr, Iran
mahdipour@kmsu.ac.ir
Morteza Khademi
Electrical Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran
khademi@um.ac.ir
Hadi Sadoghi Yazdi
Computer Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran
h-sadoghi@um.ac.ir

Received: 03/Mar/2015 Revised: 20/Jul/2016 Accepted: 25/Jul/2016

Abstract

Most popular fusion methods have their own limitations; e.g. OWA (order weighted averaging) has “linear model” and “summation of inputs proportions in fusion equal to 1” limitations. Considering all possible models for fusion, proposed fusion method involve input data confusion in fusion process to segmentation. Indeed, limitations in proposed method are determined adaptively for each input data, separately. On the other hand, land-cover segmentation using remotely sensed (RS) images is a challenging research subject; due to the fact that objects in unique land-cover often appear dissimilar in different RS images. In this paper multiple co-registered RS images are utilized to segment land-cover using FCM (fuzzy c-means). As an appropriate tool to model changes, fuzzy concept is utilized to fuse and integrate information of input images. By categorizing the ground points, it is shown in this paper for the first time, fuzzy numbers are need and more suitable than crisp ones to merge multi-images information and segmentation. Finally, FCM is applied on the fused image pixels (with fuzzy values) to obtain a single segmented image. Furthermore mathematical analysis and used proposed cost function, simulation results also show significant performance of the proposed method in terms of noise-free and fast segmentation.

Keywords: Fusion; Land-cover Segmentation; Multiple High-spatial Resolution Panchromatic Remotely Sensed (HR-PRS) Images; Fuzzy C-means (FCM).

1. Introduction

An important task in remote sensing (RS) applications is categorization of image pixels into homogeneous regions, whereas each of them corresponds to a particular land-cover type. This problem has often been modeled as **segmentation** problem and one of most utilized method to solve it, is clustering [1]. Nowadays land-cover segmentation using RS images, especially high-spatial resolution panchromatic remotely sensed (HR-PRS) ones (because of their high spatial resolution), becomes a challenging research task, due to this fact that objects in the unique scene (land-cover) often appear dissimilar in different RS images and sometimes incorrect (in someone) [2], though the land-cover has not changed. Generally obtained pixels values in RS images will be different with reality, which can be measurable by spectrometer in the ground field test, due to the following reasons:

- In the state of not being so obvious and where the (radiometric and geometric) preprocessing cannot model and remove sensor and atmosphere defects completely, the difference will be categorized as an un-sensible noise, caused by un-compensated sensor and atmosphere factors.
- On the other hand, if it happen the difference to be noticeable, as a matter of changes such as placing in shadow or white cloudy dots, it is called un-expected noise.

On the other hand, by review the related papers to RS, it is observed there are two types of uncertainties in panchromatic images. The spatial uncertainty, as the famous one, imply to this fact that there is no an exact crisp separation between various land-covers types and the segments boundaries usually express softly continuous and using fuzzy sets [3-13]. The other uncertainty in RS images is the inherent one that imply to the inaccuracy and problems in sensing and digitizing the real phenomena [7]. There are two methods parametric and deterministic to model and analysis of inherent uncertainty. In the parametric one, the uncertainty is modeled using probability density function (pdf) [14-15]. Difficulty in allocating a pdf to a pixel and un-independence of neighborhood pixels are some problems of parametric method utilizing in RS applications [10,16,17]. In the deterministic method that is used in this paper, a symbolic interval number is used (instead of crisp one) to model the inherent uncertainty [18-20].

Utilization of **multiple RS images** to measure the ground physical and geometrical properties is a conventional way to overcome the uncertainty too [21-23]. Essentially, the concern of multi-images (multi-sensors) segmentation is minimizing the uncertainty. Furthermore,

* Corresponding Author

advantage of sharing various satellites data for using their capabilities simultaneously, encourages researchers to make use of multiple RS images. In this paper **HR-PRS images** are chosen to be utilized in land-cover segmentation because of their availability for us. Although by availability of other RS images (such as multi-spectral or even hyper ones), they can be used in the same procedure to land-cover segmentation too.

On the other hand, **Data fusion** is an effective way for optimum utilization of large volume data and combines different pieces of information into some new compatible information or more accurate data [24]. Application of data fusion methods varies a lot from military applications (such as target tracking and target recognition) to non-military ones (for example machine vision, robotics and medical). Data fusion tries to perform: 1) fusion of temporal information or 2) fusion of dissimilar information and or 3) fusion of similar information from different sources (or in fixed sensing object, fusion of information obtained by one unique sensor in various conditions and times) [24]. The first two categories are examined extensively in RS applications, named as multi-temporal [25] and conventional RS images fusion (producing high-resolution multispectral images from a high-resolution panchromatic image and a low-resolution multispectral image) [3] respectively. Despite that, there have not been many works for the third category, which is the category of proposed method in this paper. **RS images fusion** can be performed at three different processing levels, according to the stage at which the fusion takes place: pixel level, feature level, and decision level [25,26]. The fusion-based multi-images segmentation methods usually perform the fusion in pixel or decision level. In the pixel level fusion-based multi-images segmentation methods that is used in this paper, by create a new fused image from input images, it is tried to improve the segmentation accuracy [25].

Since the purpose of this paper is noise-free and correct segmentation, the multi-image segmentation using fusion method (in the pixel level) is concern of this paper. Although there are many new researches in the land-cover segmentation using a single RS image (for example see [27,28]), but no new method in multi-images segmentation has been clarified. Inspecting **multi-images segmentation**, two groups of researchers published some papers [21-23,29-33]. The first group (Lee et.al. [21-23]) integrated the data from individual sensors into a set of multidimensional data to segmentation using hierarchical clustering. Against, Pieczynski et al (the second group) [29-33] used hidden Markov model (HMM) to land-cover segmentation using multi RS images.

Generally the conventional fusion methods (for example let HMM, Kalman filter, order weighted averaging (OWA) or etc.), where they are used for multi-images segmentation, let a specific proportion for each input image in the fusion process (or in the final fused image), for all pixels. While this proportion must be certificated for each pixel separately based on this fact that ambiguity in the available values of the pixel is high or low. Indeed since fuzzy concept can model existent inherent diversity and ambiguity in the available values for each land-cover point as well, the proposed method in this paper fuses the input images in one fuzzy image (an image that its pixels have fuzzy numbers instead of crisp ones). Finally the fuzzy c-means (FCM), as a commonly clustering method and a method with fuzzy output to model the spatial uncertainty, is applied on the fuzzy fused image pixels to obtain a single segmented image for each land-cover. The mathematical analysis show better performance of the proposed method in compare to the classical methods and conventional fusion methods. This is performed using a new proposed cost function. Simulation results confirm the efficiency of the proposed method in noise-free and fast segmentation aspects.

The **remainder of this paper** is organized as follows. The motivation of this paper, as a preface to understand the necessity of using non-crisp (fuzzy) numbers in land-cover segmentation using HR-PRS images is obtained in section 2. Section 3 introduces some preliminaries including the FCM clustering algorithm where it is applied on crisp and non-crisp numbers. Section 4 presents the proposed method to land-cover segmentation using multiple co-registered panchromatic images. Furthermore some analyses and comparisons are obtained in section 4 too. Simulation results are obtained in Section 5. Finally this paper is concluded in Section 6.

2. Motivation

In this section, as a preface to understand the necessity of using non-crisp (fuzzy) numbers in land-cover segmentation using RS images, land-cover points are categorized according to available multi RS images from that land-cover (for the first time). For simplicity this work is performed for images (HR-PRS ones) and based on spectral feature in this paper; however it can be performed for other signals (except image) and based on other (non-spectral) image features similarly.

Let the actual value of under-studying land-cover points (n points) in the studied feature (gray-scale level) be denoted by $\breve{X} = \{\breve{x}_1, \cdots, \breve{x}_k, \cdots, \breve{x}_n\}$, which can be achieved using spectrometer in ground field test. Considering unavailability of these values (because of un-performing ground field experiments) and the range of numbers which can be specified to each segment of land-cover (for example let a river as a segment that its pixels values vary with depth of water, water impurity, kind of river bottom and etc.), an interval is considered for the actual value of each pixel ($\breve{x}_k \in \left[\breve{x}_k^{min}, \breve{x}_k^{max}\right], k = 1, \cdots, n$).

Letting L available (co-registered) HR-PRS images ($X^{(l)} = \left\{x_1^{(l)}, \cdots, x_k^{(l)}, \cdots, x_n^{(l)}\right\}$, $l = 1, \cdots, L$), It is observed that, nevertheless the covered land-cover by all L images and even the used sensor for imaging may be same, but most corresponding pixels in L images have different values. This phenomenon can be seen easily in the first scene images and Figure 1 that are obtained using IRS-P5 satellite. Comparing values of first scene images, pixel by pixel, it can be observed that only 0.84% of pixels have same values in both images, for 62.54% of pixels, image A and for the remainder (36.61%), image B has more values (e.g. under shadow region of image A, mentioned by circle). Therefore, the minimum and maximum of observed values sets can be defined as follows:

$$\mathrm{X}^{\min} = \left\{x_1^{\min}, \cdots, x_n^{\min}\right\} \text{ where } x_k^{\min} = \min\left\{x_k^{(1)}, \cdots, x_k^{(L)}\right\} \text{for } k = 1, \cdots, n \qquad (1)$$

$$\mathrm{X}^{\max} = \{x_1^{\max}, \cdots, x_n^{\max}\} \text{ where } x_k^{\max} = \max\left\{x_k^{(1)}, \cdots, x_k^{(L)}\right\} \text{for } k = 1, \cdots, n \qquad (2)$$

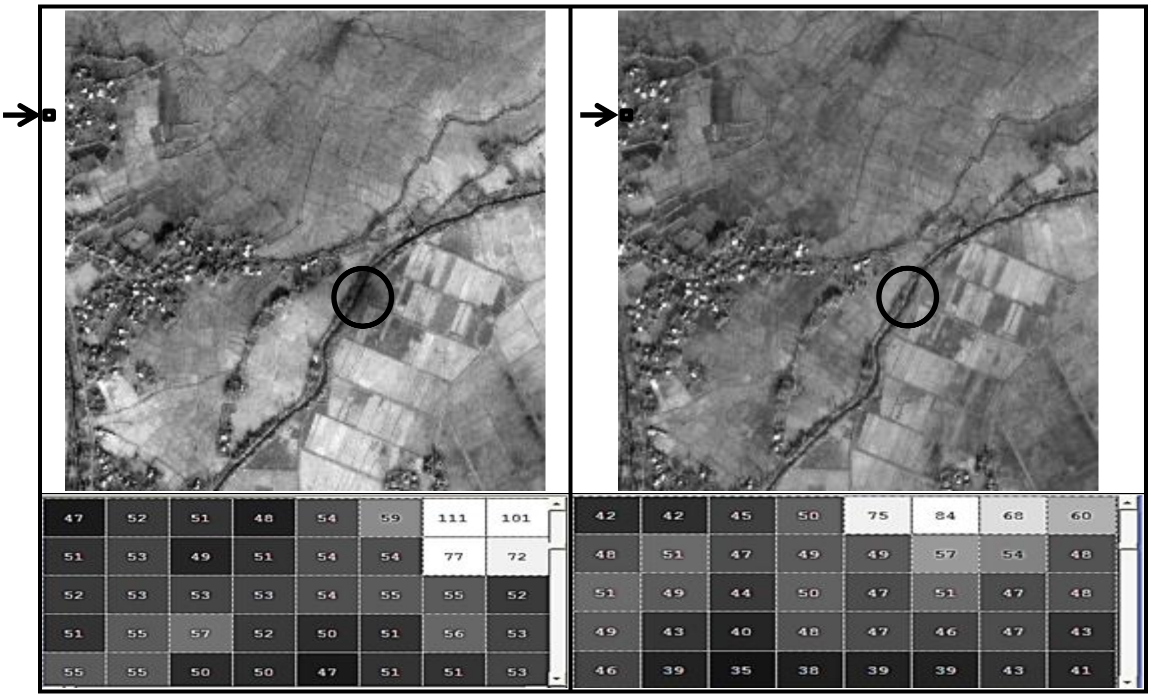
According to above notes and also considering the variation of $x_k^{(l)}$ and $\check{x}_k$ ($x_k^{(l)} \in \left[x_k^{\min}, x_k^{\max}\right]$ and $\check{x}_k \in \left[\check{x}_k^{\min}, \check{x}_k^{\max}\right]$) each point of land-cover will belong to one of two following categories based on the relation between actual values range ($\left[\check{x}_k^{\min}, \check{x}_k^{\max}\right]$) and range of obtained values by L images ($\left[x_k^{\min}, x_k^{\max}\right]$).

- First category land-cover points are not influenced by unwanted and unexpected happenings (such as placing under white cloudy dot or in the shadow region) in the input images. Since the (radiometric and geometric) preprocessing cannot model and remove sensor and atmosphere defects completely, this category is affected only by un-modeled (or more precisely, weak modeled) causes of sensor and atmosphere, named as un-sensible noise in this paper. The considered model for atmosphere, compose of various probability distributions (for example see [34]), is an acceptable reason for this fact that atmosphere defect cannot be removed from RS images completely. Therefor the relations between variation ranges of $x_k^{(l)}$and $\check{x}_k$ for each point of this category will be one of nine states Figure 2-A. For this group of pixels, the actual value $\check{x}_k$ ($\check{x}_k \in \left[\check{x}_k^{\min}, \check{x}_k^{\max}\right]$ is shown by bilateral black vector in Figure 2-A) is influenced by un-sensible noise. So the variation range of obtained values by input images ($x_k^{(l)} \in \left[x_k^{\min}, x_k^{\max}\right]$ is shown by one sided blue vector in Figure 2-A) will be small and in the order of $\check{x}_k$variation range.
- Second group points are affected by un-wanted (un-expected) noise (e.g. placing in shadow or white cloudy dots) in addition un-sensible one, at least in one input image. This group of pixels is supposed to be affected by un-sensible noise as the simplest state (top-left state of Figure 2-A). Major reason of this assumption is more role of un-expected noise comparing the un-sensible one in the values of these pixels. The relation between the variations ranges of $x_k^{(l)}$ and $\check{x}_k$ in this pixels categorization will be accessible through one of following three states, Figure 2-B:D.

According to the mentioned notes, a segmentation methodology must be selected that obtain noise-free labels (segments) for all land-cover points using multi-images, of which the pixels would have each states of Figure 2-A:D.

In the other view, according to relations (1) and (2), an n-dimensional hyper-cube can be defined as the input space in segmentation of land-cover using multi-images. The range of each dimension (for example the dimension k-th) is from $x_k^{\min}$ to $x_k^{\max}$. For $n = 2$ this input space is illustrated in Figure 2-E.

According to above explanations (performed categorization) it can be concluded that it is very probable (except in $\check{x}_k \notin \left[x_k^{\min}, x_k^{\max}\right]$ cases) to present the actual values of land-cover points ($\check{\mathrm{X}} = \{\check{x}_1, \cdots, \check{x}_k, \cdots, \check{x}_n\}$) by a point in this hyper-cube (inside it or on its boundary). This point which can be named as the **optimum point in the input space (OPIS)** is unknown when only RS images are used to segment and any ground truth data is not available. Obviously, applying the clustering method (e.g. FCM) on the OPIS, the optimum response (optimum centers and membership values and finally optimum segmented image) can be reached.



47	52	51	48	54	59	111	101
51	53	49	51	54	54	77	72
52	53	53	53	54	55	55	52
51	55	57	52	50	51	56	53
55	55	50	50	47	51	51	53

42	42	45	50	75	84	68	60
48	51	47	49	49	57	54	48
51	49	44	50	47	51	47	48
49	43	40	48	47	46	47	43
46	39	35	38	39	39	43	41

Fig. 1. The first land-cover images (left: A and right: B) and corresponding values of mentioned regions by →▪.

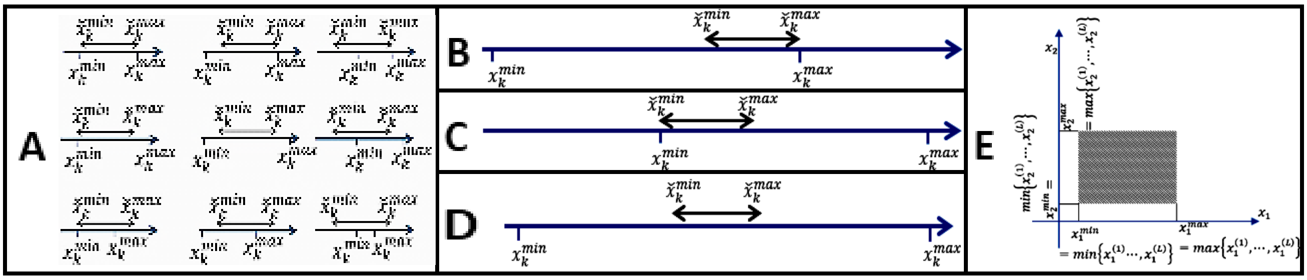

A
B
C
D
E
$\breve{x}_k^{min}$
$\breve{x}_k^{max}$
x_k^{min}
x_k^{max}
x_1
x_2
$x_1^{min} = min\{x_1^{(1)},\cdots,x_1^{(L)}\}$
$x_1^{max} = max\{x_1^{(1)},\cdots,x_1^{(L)}\}$
$x_2^{min} = min\{x_2^{(1)},\cdots,x_2^{(L)}\}$
$x_2^{max} = max\{x_2^{(1)},\cdots,x_2^{(L)}\}$

Fig. 2. The corresponding figures of land-cover points categorizing; the variation ranges of actual pixel number ($\breve{x}_k \in [\breve{x}_k^{min}, \breve{x}_k^{max}]$ is shown by bilateral vector) and the obtained values by L input images for k-th pixel ($x_k^{(l)} \in [x_k^{min}, x_k^{max}]$ is shown by one sided vector), where in A it is influenced only by un-sensible noise in all L input images; in B it only is placed in shadow at least in one input image, in C it only is placed under white cloudy dot at least in one input image, in D it is placed in shadow and white cloudy dot at least in two separate input images; E. The input space of multi-images segmentation (hyper-cube for $n = 2$)

3. Preliminaries

In this section, as the proposed method is based on the conventional FCM, a description of FCM will be given concerning the special manner in which it is applied on crisp numbers. Then symbolic interval numbers as the simplest fuzzy numbers type along with the FCM applying on them, are explained.

3.1 FCM Algorithm and Crisp Numbers

The fuzzy c-means (FCM) algorithm, the best-known clustering algorithm, has been used in a wide range of engineering and scientific disciplines such as medicine imaging, bioinformatics, pattern recognition, and data mining [35,36]. FCM clustering method assigns fuzzy memberships to each input member. This method is fuzzy equivalence of the nearest center “hard” clustering method. The aim of FCM algorithm is minimizing the following objective function ($J(U,V)$) with respect to fuzzy memberships $U = [u_{i,k}]_{c\times n}$ and cluster centers $V = \{v_1,\cdots,v_i,\cdots,v_c\}$,

$$J(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{i,k}^{\theta} d^2(x_k, v_i) \qquad (3)$$

$$d^2(x_k, v_i) = (x_k - v_i)^T (x_k - v_i) = \|x_k - v_i\|^2 \qquad (4)$$

Where n is the number of input numbers, c is the number of clusters, $X = \{x_1,\cdots,x_k,\cdots,x_n\}$ is the set of input numbers which is a finite set of p-dimensional vectors on the real numbers ($x_k = [x_{k,1},\cdots,x_{k,p}]^T \in \mathbb{R}^p$ for $k = 1,\cdots,n$) and $\theta > 1$ is the fuzziness index. The matrix $U = [u_{i,k}]_{c\times n}$ is called the fuzzy membership degree with following constraint:

$$\begin{cases} u_{i,k} \in [0,1]\,, \text{for } i = 1,\cdots,c \ \text{ and } \ k = 1,\cdots,n \\ \sum_{i=1}^{c} u_{i,k} = 1\,, \ k = 1,\cdots,n \end{cases} \qquad (5)$$

Where $u_{i,k}$ is the membership grade of k-th input number to i-th cluster. $V = \{v_1, \cdots, v_i, \cdots, v_c\}$ is the cluster prototypes (centers) set, $v_i = [v_{i,1}, \cdots, v_{i,p}]^T \in \mathbb{R}^p$ for $i = 1, \cdots, c$ is the center of i-th cluster and $d^2(x_k, v_i)$ denotes the Euclidean distance of x_k and v_i in p-dimensions space.

Creating Lagrange function $L(U, V, \lambda)$ and using Lagrange multiplayers $\lambda_k, k = 1, \cdots, n$, the objective function $J(U, V)$ can be minimized subject to constraints (5) to conclude updating relations as follows:

$$L(U, V, \lambda) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{i,k}^{\theta} d^2(x_k, v_i) - \sum_{k=1}^{n} \lambda_k \left(\sum_{i=1}^{c} u_{i,k} - 1\right) \quad (6)$$

$$\partial L(U, V, \lambda)/\partial u_{i,k} = 0 \Rightarrow \cdots \Rightarrow u_{i,k} = \left(\sum_{r=1}^{c} \left(\frac{d^2(x_k, v_i)}{d^2(x_k, v_r)}\right)^{\frac{1}{\theta-1}}\right)^{-1} \quad (7)$$

$$\partial L(U, V, \lambda)/\partial v_i = 0 \Rightarrow \cdots \Rightarrow v_i = \frac{\sum_{k=1}^{n} u_{i,k}^{\theta} x_k}{\sum_{k=1}^{n} u_{i,k}^{\theta}} \quad (8)$$

where $d^2(x_k, v_i) = (x_k - v_i)^T (x_k - v_i)$ is the euclidean distance of x_k and v_i and $i = 1, \cdots, c$ and $k = 1, \cdots, n$.

3.2 FCM Algorithm and Non-Crisp Numbers

Since for all studied scenes (land-covers) in this paper only two HR-PRS images are available and symbolic interval numbers are the best fuzzy (non-crisp) numbers to represent two available values of a fact (in this paper, each land-cover point), in this sub-section the symbolic interval numbers as the simplest non-crisp (fuzzy) numbers are introduced and discussed. A non-crisp number $\tilde{x}$ will be supposed as a symbolic interval number (SIN) if its membership function be expressed as follows [37]:

$$\mu_{\tilde{x}}(x) = \begin{cases} 1 \ , \ \alpha_{\tilde{x}} \le x \le \beta_{\tilde{x}} \\ 0 \ , \ \text{Otherwise} \end{cases} \quad (9)$$

A SIN $\tilde{x}$ can be denoted with its start point ($\alpha_{\tilde{x}}$) and its end point ($\beta_{\tilde{x}}$) as $\tilde{x} = (\alpha_{\tilde{x}}, \beta_{\tilde{x}})_{SIN}$.

Suppose two symbolic interval numbers $\tilde{x}_k$ and $\tilde{v}_i$ in p dimensions space, $\tilde{x}_k = \{\tilde{x}_{k,1}, \tilde{x}_{k,2}, \cdots, \tilde{x}_{k,p}\}$, $\tilde{x}_{k,j} = \left(\alpha_{\tilde{x}_{k,j}}, \beta_{\tilde{x}_{k,j}}\right)_{SIN}$ and $\tilde{v}_i = \{\tilde{v}_{i,1}, \tilde{v}_{i,2}, \cdots, \tilde{v}_{i,p}\}$, $\tilde{v}_{i,j} = \left(\alpha_{\tilde{v}_{i,j}}, \beta_{\tilde{v}_{i,j}}\right)_{SIN}$ for $j = 1, 2, \cdots, p$. The used metric (dissimilarity or distance) for SINs in [37], as the simplest metric in the symbolic interval numbers case which is used in this paper, is as follows:

$$d^2(\tilde{x}_k, \tilde{v}_i) = \sum_{j=1}^{p} \left(\left(\alpha_{\tilde{x}_{k,j}} - \alpha_{\tilde{v}_{i,j}}\right)^2 + \left(\beta_{\tilde{x}_{k,j}} - \beta_{\tilde{v}_{i,j}}\right)^2\right) \quad (10)$$

Based on above metric, the FCM can be applied on SINs as follows [37]:

$$\underset{\tilde{V}, U}{\text{argmin}} \left\{J(\tilde{V}, U) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{i,k}^{\theta}\, d^2(\tilde{x}_k, \tilde{v}_i)\right\} \Rightarrow \cdots \Rightarrow$$

$$\text{subject to: } \sum_{i=1}^{c} u_{i,k} = 1 \ , \ \text{for } k = 1, \cdots, n$$

$$\begin{cases} u_{i,k} = \left(\sum_{j=1}^{c} \left(\frac{d^2(\tilde{x}_k, \tilde{v}_i)}{d^2(\tilde{x}_k, \tilde{v}_i)}\right)^{\frac{1}{\theta-1}}\right)^{-1} \\ \alpha_{\tilde{v}_{i,j}} = \frac{\sum_{k=1}^{n} u_{i,k}^{\theta} \alpha_{\tilde{x}_{k,j}}}{\sum_{k=1}^{n} u_{i,k}^{\theta}} \\ \beta_{\tilde{v}_{i,j}} = \frac{\sum_{k=1}^{n} u_{i,k}^{\theta} \beta_{\tilde{x}_{k,j}}}{\sum_{k=1}^{n} u_{i,k}^{\theta}} \end{cases} \quad (11)$$

Having fulfilled the preliminaries, the proposed method will be presented in the next section.

4. Proposed Method

This section of paper proposes a land-cover segmentation algorithm using available multi HR-PRS images from the land-cover and compares it analytically with other methods in the following sub-sections.

4.1 Descriptions

The structure of the proposed method, containing 5 blocks, is as Figure 3 which will be explained consequently block by block.

HR-PRS input images: As mentioned, L HR-PRS images with same spatial-resolution are considered as input images. These images, registered together and denoted by $X^{(l)} = \left\{x_1^{(l)}, \cdots, x_n^{(l)}\right\}, l = 1, \cdots, L$ belong to same unchanged scene.

Pixels fuzzifying: As mentioned in introduction, the nature of proposed method and the considered problem in this paper differs with normal concept of RS images fusion [3]. But in the other view, the proposed method can be considered as a method to fuse multi HR-PRS images and applying the segmentation process on the fused image. On the other hand, fuzzy concept, in comparison with crisp one and in the way it was described, can represent land-cover points completely when multi images are available. Therefor the proposed method considers a fuzzy (non-crisp) number value for each fused image pixel. This fuzzy number is obtained using L available values for each land-cover point. The fuzzy number can be a symbolic interval type fuzzy one in the simplest state. LR-Type, Gaussian (normal), triangular, trapezoidal and etc. are other types of fuzzy numbers that can be used. Since, in this paper only two images are available for each land-cover (L = 2), based on two available values for each land-cover point, the small one is considered as x_k^{min} and the other one as x_k^{max}. Finally, each pixel in the fused image is demonstrated by $\tilde{x}_k = \tilde{x}_{k,SIN} = \left(\alpha_{\tilde{x}_k}, \beta_{\tilde{x}_k}\right)_{SIN} = \left(x_k^{min}, x_k^{max}\right)_{SIN}$, indeed, $\alpha_{\tilde{x}_k} = x_k^{min}$ and $\beta_{\tilde{x}_k} = x_k^{max}$.

FCM and Determine optimum clusters number: let’s suppose that we want to cluster the land-cover (fused image) to c (c ≥ 2) clusters. The proposed method applies FCM on the fused image, just one time and according to the mentioned procedure in section (3.2), to obtain a single segmented image for each land-cover. As far as a mathematical based fusion concerns, the proposed method can be presented as follows:

$$\tilde{x}_k = \mathrm{Fusion}\left(x_k^{(1)}, \cdots, x_k^{(l)}, \cdots, x_k^{(L)}\right) \Longrightarrow$$
$$\underset{\tilde{V}, U}{\operatorname{argmin}} \left\{J(\tilde{V}, U) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{i,k}^{\theta}\, d^2(\tilde{x}_k, \tilde{v}_i)\right\} \tag{12}$$
$$\text{subject to: } \sum_{i=1}^{c} u_{i,k} = 1 \ \ , \ \text{for } k = 1, \cdots, n$$

Where, the notation $\sim$ is used to represent fuzzy (in this section, symbolic interval) numbers, n is the number of per input image pixels (or under-studying land-cover points), c is the number of clusters, $\tilde{V} = \{\tilde{v}_1, \cdots, \tilde{v}_c\}$ and $\tilde{v}_i$ is the center of i-th cluster, $d(,)$ represents the distance of two SINs according to (10), $\theta > 1$ is the fuzziness index and the matrix $U = [u_{i,k}]_{c\times n}$ is called the fuzzy membership degree.

It must be mentioned, one of problems in image segmentation is obtaining optimum number of segments for each input land-cover [38]. In the clustering based segmentation methods, this is equivalent to optimum clusters number determination process (named OCNDP in this paper) for the fused image usually performed by optimizing a cluster validity index where varies with the number of clusters [39]. This paper uses Xie–Beni (XB) cluster validity index [40] for finding the optimum number of clusters. The XB index is defined as a function of total variation ratio to clusters’ centers minimum separation as follows:

$$XB(c) = J/\left(\min_{i\neq j}\{d^2(v_i, v_j)\}\right) \tag{13}$$

Where, J is the minimized objective function by clustering algorithm, and v_i and v_j are two separate clusters centers.

Since the centers in the proposed method are non-crisp, v_i and v_j are replaced by $\tilde{v}_i$ and $\tilde{v}_j$ respectively in equation (13) and $d^2(\tilde{v}_i, \tilde{v}_j)$ is obtained using equation (10). Therefore, the optimum number of clusters $\hat{c}$ can be achieved for each land-cover using the following equation:

$$\hat{c} = \underset{c}{\operatorname{argmin}} XB(c) \tag{14}$$

As a result, outcomes of FCM applied on the fused image (where the number of clusters is $\hat{c}$) are considered to be the output of this block and consequently, input for the next block (Defuzzification).

Defuzzification: Having clustered fused image (to $\hat{c}$ clusters), to assign the pixels to correspondent segment, it is decided according to maximum membership value. Each pixel will be assigned to the cluster containing the highest membership value comparing other clusters. Hence by sorting the clusters based on their centers absolute values in an ascending order from 1 to $\hat{c}$ and labeling the pixels, the **segmented image** for each land-cover will be resulted.

4.2 Analysis and Comparisons with other Segmentation Methods

Concerning input and output space (Figure 2.E), the proposed method selects all input space points and applies FCM (and consequently performing OCNDP) on one shot. Eventually, the optimum result will be among the obtained answers and can be achieved using a suitable metric and defuzzification process. Interestingly, achieving this important advantage, the computational complexity in this method would be to the order of classical methods [27,28] (one time using from clustering method and OCNDP). Classical methods [27,28] select a single point from the input space (which is equivalent to a single input image) and apply clustering method and OCNDP on that point to obtain land-cover segmentation results. It is seen the considered point from input space in these methods (and consequently the resulted outcomes) is absolutely different with the OPIS. Consequently, the obtained segmentation results would similarly differ with the actual ones (obtained results by ground experiments). From this point of view, the proposed method is a highly extended form of the classical methods. It means, the number of selected points from input space to segmentation, denoted by Ĺ, in the classical methods is limited (Ĺ = 1), but in the proposed method is unlimited (Ĺ → +∞). Therefore, the proposed method no have classical methods defect, high distance between the optimum and resulted answer.

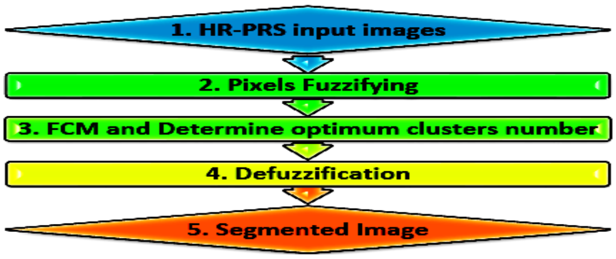


Fig. 3.The block diagram of the proposed method.

Comparing the proposed method with other multi-images segmentation methods [21-23,29-33], they apply clustering on L images or L points of input space (Ĺ = L) and result output which is obtained by fusing these clustering outputs with the help of conventional fusion methods (perform fusing in the decision stage by using a known and conventional method e.g. Dempster- Shafer evidence theory). In addition to the mentioned defect of classical methods, the ambiguity in the number of optimum clusters determination (because, each input image may obtain different number in OCNDP), the way to fuse the L segmentation results and very high computational complexity (because of two reasons: 1. L times performing OCNDP applied on L input images separately while this action is time-consuming even for one time or one image; 2. Applying clustering method on

L input images) are other defects of these methods. Finally, we can summarize these comparisons as Table 1.

Eventually, it can be listed the advantages of the proposed method as follows:

1. Low computational complexity and avoiding confusion in OCNDP performing.
2. Considering OPIS as the input for segmentation and consequently obtaining the optimum response.
3. Reduction of pixels values variation effects on the segmentation results (because of the fuzzy clustering method's robustness versus the noise [41]).
4. Using various existent metrics for symbolic interval numbers [37,42-44] in the FCM applying and obtaining various correspondent responses, where each of them have their own different properties.

Capability of detecting the pixels containing high uncertainty (the land-cover points with high available values ranges). This capability can provide supervised decision about them.

4.3 New Interpretation and Comparison with Conventional Fusion Methods

In this subsection, the proposed method will compare in details with the known conventional fusion methods by mathematical analysis, where they fuse the input images and result a crisp fused image. Then the conventional FCM is applied on the resulted fused image. It must be mentioned, for simplicity, in this sub-section it is supposed $\mathrm{L} = 2$.

In continue, firstly the performances of the conventional min, max, mean, median, OWA and Kalman-filter fusion methods are analyzed in the input space Figure 2-E, for $\mathrm{L} = 2$. Then by proposing suitable cost functions, it will be shown that the proposed method have better performance in compare to them.

The min, max, mean and median operators (or fusion methods) have fix performance and result a fix point from the input space in the presented problem statement (input space Figure 2-E, for $\mathrm{L} = 2$). While the location of fusion results in these methods are illustrated in Figure 4, their performances can be expressed mathematically as follows:

$$\mathrm{x_k} = \mathrm{Fusion}\left(\mathrm{x_k^{(1)}}, \mathrm{x_k^{(2)}}\right) = \mathrm{Fusion}\left(\mathrm{x_k^{min}}, \mathrm{x_k^{max}}\right) = \begin{cases} \mathrm{x_k^{min}} \ , \ \text{if Fusion: Min} \\ \mathrm{x_k^{max}} \ , \ \text{if Fusion: Max} \\ \left(\mathrm{x_k^{min}} + \mathrm{x_k^{max}}\right)/2 \ , \text{if Fusion: Mean or Median} \end{cases} \quad (15)$$

The OWA fusion method is a general form of examined methods min, max, mean and median. The OWA performance in the considered problem can be represented as follows:

$$\mathrm{x_k} = \mathrm{Fusion}\left(\mathrm{x_k^{(1)}}, \mathrm{x_k^{(2)}}\right) = \mathrm{w}\mathrm{x_k^{(1)}} + (1-\mathrm{w})\mathrm{x_k^{(2)}} \ , \ 0 \le \mathrm{w} \le 1 \ , \ \text{if Fusion: OWA} \quad (16)$$

where the weight value w usually is satisfied by optimizing an specific cost function. For example in the availability of ground truth data $\breve{\mathrm{X}} = \{\breve{\mathrm{x}}_1, \cdots, \breve{\mathrm{x}}_\mathrm{k}, \cdots, \breve{\mathrm{x}}_\mathrm{n}\}$, the considered objective function for optimizing can be the mean square error between the fused values $\{\mathrm{x_k}\}_{\mathrm{k}=1}^{\mathrm{n}}$ and the actual ones $\{\breve{\mathrm{x}}_\mathrm{k}\}_{\mathrm{k}=1}^{\mathrm{n}}$ as follows that must be minimized.

$$\mathrm{CF} = \sum_{\mathrm{k}=1}^{\mathrm{n}} (\mathrm{x_k} - \breve{\mathrm{x}}_\mathrm{k})^2 \quad (17)$$

Indeed the OWA checks the CF value for all allowed w values ($0 \le \mathrm{w} \le 1$) and selects the best w value, where the CF will be optimized. This concept is equivalent to checking all points on the main diagonal of input space rectangle, point by point, and selecting the best one (as illustrated in Figure 4).

In the last case, the Kalman filter fusion method is examining. This method is as same as the OWA method with this difference that an additive white Gaussian noise has been added. The corresponding relation for this fusion method in the examining problem state is as follows:

$$\mathrm{x_k} = \mathrm{Fusion}\left(\mathrm{x_k^{(1)}}, \mathrm{x_k^{(2)}}\right) = \mathrm{w}\mathrm{x_k^{(1)}} + (1-\mathrm{w})\mathrm{x_k^{(2)}} + \mathrm{n_k}, 0 \le \mathrm{w} \le 1 \ , \mathrm{n_k} \sim \mathrm{N}(0, \sigma^2), \text{if Fusion: Kalman} \quad (18)$$

The resulted point set using the equation (18) is illustrated in Figure 4. It must be mentioned, the wideness of searching region (region denoted by ▨ in Figure 4) is related to the variance of the additive noise. By increasing the variance, the wideness region will grow up and may reach to the better response; against the probability of reaching to the best point in searching region (the nearest point of search region to the OPIS) will be decreased and it will be a time-consuming process and vice versa.

Generally the set of mentioned points from the input space by these fusion methods (represented by $\mathrm{I_{method}}$) can be represented by equations of (19).

It is observed, the conventional fusion methods consider a fix point (min, max, mean and median methods) or search within points on the line (the OWA method) or its around (the Kalman filter method), point by point, to get the best point in the search region. While the OPIS is not located in the search region necessarily.

Against, the proposed method consider all input space points one shot (not by search and testing point by point, as same as the examined methods) and results the best point from input space (nearest point to OPIS) as fusion method output. Therefor the set of mentioned points from the input space by proposed method can be represented as follows.

$$\mathrm{I_{Proposed}} = \mathrm{I} = \left\{\mathrm{x_k} \ , \text{for } \mathrm{k} = 1, \cdots, \mathrm{n} \middle| \mathrm{x_k^{min}} \le \mathrm{x_k} \le \mathrm{x_k^{max}}\right\} (20)$$

Comparing the mentioned points sets of introduced methods, the following equation can be concluded.

$$\mathrm{I_{min}}, \mathrm{I_{max}}, \mathrm{I_{mean}}, \mathrm{I_{median}} \subseteq \mathrm{I_{OWA}} \subseteq \mathrm{I_{Kalman}} \subseteq \mathrm{I_{Proposed}} = \mathrm{I} \quad (21)$$

In order to compare the various examined fusion methods, supposing availability of OPIS (ground truth data), the minimum distance achievable by each fusion method (by performing full search in OWA and Kalman) is proposed as an objective function in this paper as follows:

$$\mathrm{Q_{e,method}} = \min\left(\sum_{\mathrm{k}=1}^{\mathrm{n}} \left\| \mathrm{Fusion}\left(\mathrm{x_k^{(1)}}, \cdots, \mathrm{x_k^{(L)}}\right) - \breve{\mathrm{x}}_\mathrm{k} \right\|^2\right) , \text{where } \mathrm{Fusion}\left(\mathrm{x_k^{(1)}}, \cdots, \mathrm{x_k^{(L)}}\right) \in \mathrm{I_{method}} \quad (22)$$

Where 'method' can be referred to each examined method.

Table 1. Summarizing comparisons of the proposed method with other land-cover segmentation methods [21-23, 27-33]

Property Method	Number of considered points from input space as input (L)	Number of performing clustering method and OCNDP	Existing the ambiguity in OCNDP for each land cover
Proposed method	$+\infty$ (contain the optimum point in input space)	1	No
other multi-images segmentation methods [21-23, 29-33]	L	L	Yes
Classical methods [27, 28]	1	1	No

$$\begin{cases} I_{min} = \left\{x_k \,, for\ k = 1,\cdots,n \middle| x_k^{min} \le x_k \le x_k^{max} \,, x_k = x_k^{min}\right\} \\ I_{max} = \left\{x_k \,, for\ k = 1,\cdots,n \middle| x_k^{min} \le x_k \le x_k^{max} \,, x_k = x_k^{max}\right\} \\ I_{mean} = \left\{x_k \,, for\ k = 1,\cdots,n \middle| x_k^{min} \le x_k \le x_k^{max} \,, x_k = \left(x_k^{(1)} + \cdots + x_k^{(L)}\right)/L\right\} \\ I_{median} = \left\{x_k \,, for\ k = 1,\cdots,n \middle| x_k^{min} \le x_k \le x_k^{max} \,, x_k = median\left(x_k^{(1)},\cdots,x_k^{(L)}\right)\right\} \\ I_{OWA} = \left\{x_k \,, for\ k = 1,\cdots,n \middle| x_k^{min} \le x_k \le x_k^{max} \,, x_k = w_1 x_k^{(1)} + \cdots + w_L x_k^{(L)} \,, w_1 + \cdots + w_L = 1\right\} \\ I_{Kalman} = \{x_k \,, for\ k = 1,\cdots,n |\, x_k^{min} \le x_k \le x_k^{max} \,, x_k = w_1 x_k^{(1)} + \cdots + w_L x_k^{(L)} + n_k \,, w_1 + \cdots + w_L = 1 \,, n_k \sim N(0,\sigma^2)\} \end{cases} \quad (19)$$

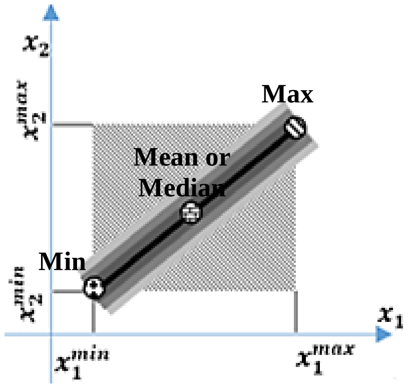


Fig. 4. The input space of multi-images segmentation and proposed method (▣); resulted points by fusion methods min (●), mean (●),median (●) and max (●); resulted possible regions by fusion methods OWA (▮) and Kalman (▥).

Computing equation (22), independent from the exact location of X̌, according to (21), following relation can be concluded:

$$Q_{e,Proposed} \le Q_{e,Kalman} \le Q_{e,OWA} \le Q_{e,min}, Q_{e,max}, Q_{e,mean}, Q_{e,median} \quad (23)$$

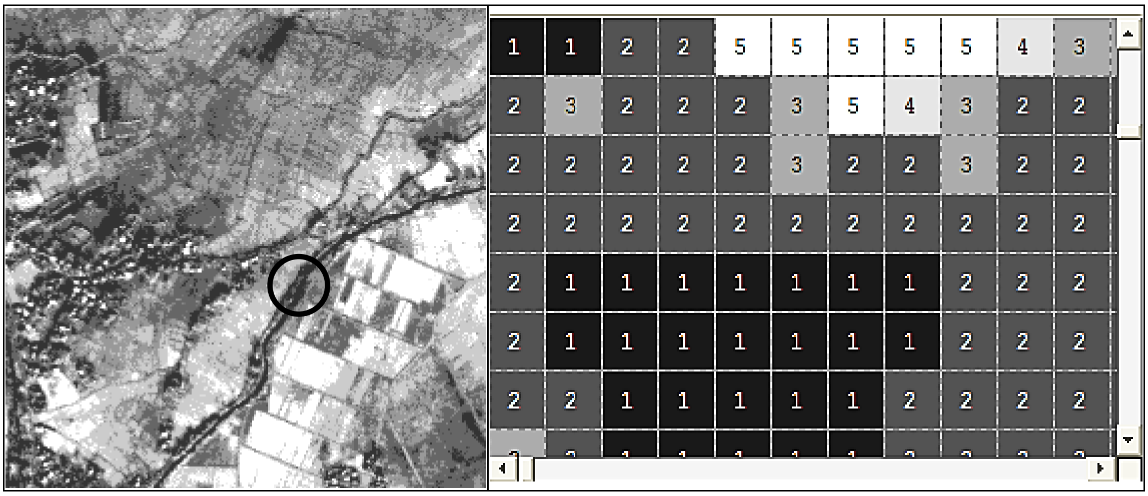
It has been observed that the fusion result point in the proposed method is nearest to OPIS in compare to other fusion methods.

5. Simulations Results

The simulations are performed in two cases and results are reported in this section. Simulations are performed on a PC equipped with an Intel(R) core(TM) i7 cpu 960 @ 3.20GHz processor, 6GB-RAM and using Win-7 64-bits and Matlab 2012-a.

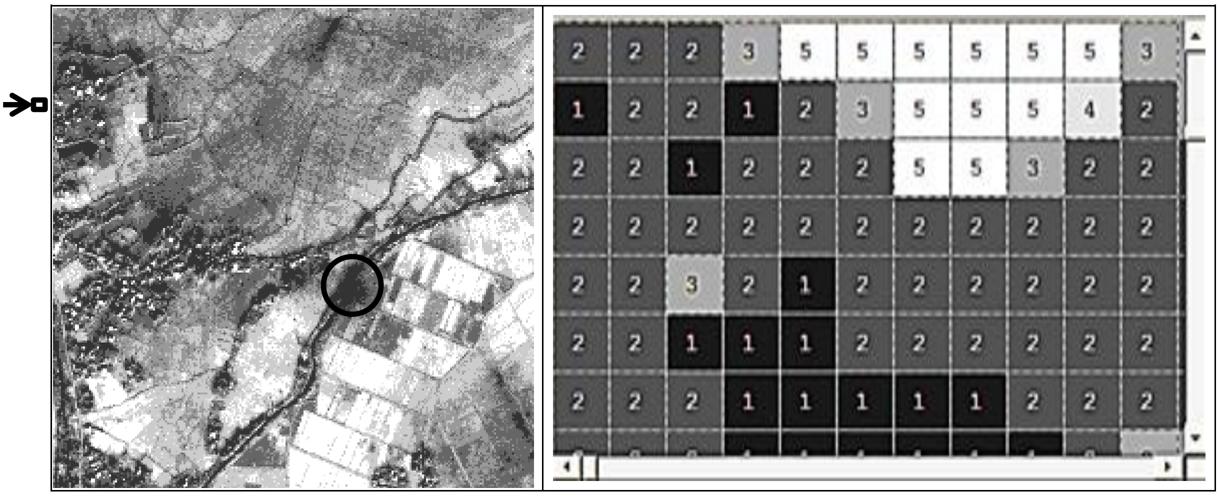
Case 1. In the first case, to subjective evaluation of the proposed method, it is applied on the images of Figure 1. Let we want to segment the corresponding land-cover of Figure 1 to 5 segments. The resulted image and corresponding labels (of mentioned region by) using proposed method is illustrated in Figure 5. To ➔■mpare it with conventional FCM, FCM is applied on the both images of Figure 1 too. Resulted images and labels of mentioned region by in this case are as Figure 6 and Figure 7. It is observed➔■while shadow effects the resulted image by FCM in the first image (mentioned by circle), but in the proposed method result and FCM applied on the second image, shadow effect is not noticeable. Furthermore, while the proposed method obtain only one label for each point, FCM generally obtain 2 (L) various labels for each point. Generally, independent from the number of land-cover images or L, only one label is derived using proposed method for each pixel, as same as the conventional fusion ones. Against, classical methods may outcome various labels for a unique scene pixels generally. This event causes confusion in the segmentation by FCM (generally, for classical methods case).

After this subjective evaluation, in continue, the proposed method will be compared with some classical fusion based methods, such as, min, mean and max in the next section.



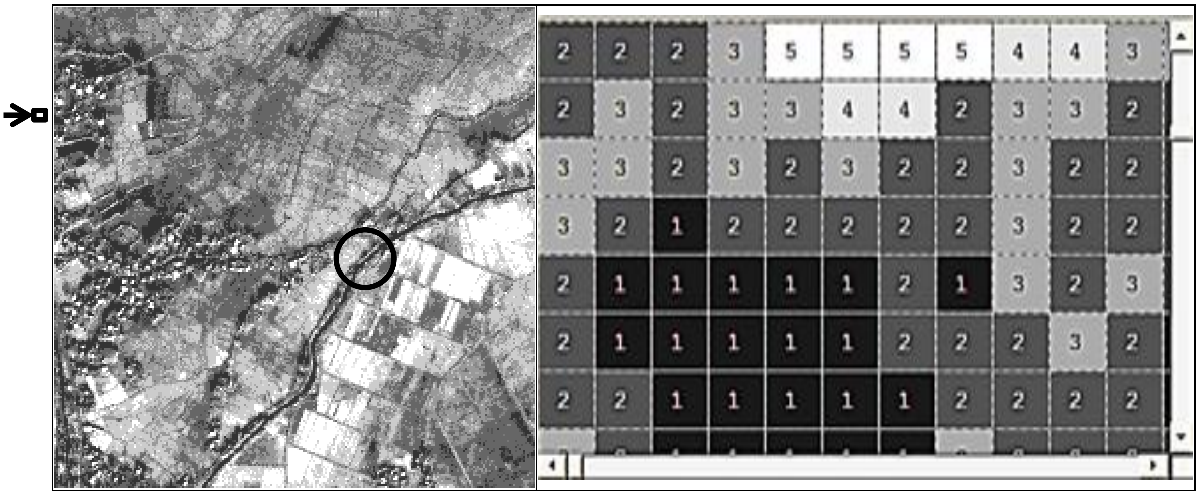
1	1	2	2	5	5	5	5	5	4	3
2	3	2	2	2	3	5	4	3	2	2
2	2	2	2	2	3	2	2	3	2	2
2	2	2	2	2	2	2	2	2	2	2
2	1	1	1	1	1	1	1	2	2	2
2	1	1	1	1	1	1	1	2	2	2
2	2	1	1	1	1	1	2	2	2	2

Fig. 5. The resulted image and corresponding labels of mentioned region by ➔□ in segmenting the first land-cover (Figure 1) using the proposed method.



2	2	2	3	5	5	5	5	5	5	3
1	2	2	1	2	3	5	5	5	4	2
2	2	1	2	2	2	5	5	3	2	2
2	2	2	2	2	2	2	2	2	2	2
2	2	3	2	1	2	2	2	2	2	2
2	2	1	1	1	2	2	2	2	2	2
2	2	2	1	1	1	1	1	2	2	2

Fig. 6. The resulted image and corresponding labels of mentioned region by ➔□ in segmenting the first image of Figure 1 using the FCM.



2	2	2	3	5	5	5	5	4	4	3
2	3	2	3	3	4	4	2	3	3	2
3	3	2	3	2	3	2	2	3	2	2
3	2	1	2	2	2	2	2	3	2	2
2	1	1	1	1	1	2	1	3	2	3
2	1	1	1	1	1	2	2	2	3	2
2	2	1	1	1	1	1	2	2	2	2

Fig. 7. The resulted image and corresponding labels of mentioned region by ➔□ in segmenting the second image of Figure 1 using the FCM.

Case 2. In this case to objective evaluation of proposed method, it is applied on the second land-cover images that are illustrated in Figure 8. This scene is located in south of Iran, with geographical location of 26.67-26.8 N and 56.04-56.06 E. Two corresponding images are obtained using Geoey-1 (0.5 m spatial resolution) and IRS-P5 (2.5 m spatial resolution) panchromatic sensors. These images are resampled and registered together and have same size (1000*1000 pixels). In order to obtain Ground truth (GT) data for accuracy assessment, a planimetric map of this area is utilized. This map is illustrated in Figure 9-A, B.

As it can be seen, the map shows exact location and structure of roads, buildings, sea, seaside and etc. While in accuracy assessment process, label of each point is needed. Therefore by using this map, on the first image of land-cover where it seems be more accurate, 4 classes (sea with black DN, main road with new and dark asphalt, sidetrack with old and bright asphalt and building's roof with white DN) are selected and labeled to 1, 2, 3 and 4 respectively, according to DN increasing manner. These 4 classes that are assumed as GT data, illustrated in Figure 9-C, D. By obtaining the GT, in continue the proposed method will be compared objectively with classical fusion based segmentation methods. In the classical fusion based segmentation methods, named as CFSM-1, CFSM-2 and CFSM-3, the input images are fused using min, mean and max operators respectively. Then FCM is applied on the fused image, to get the segmented image. The performance of these 4 methods (proposed and 3 classical fusion based segmentation ones) are evaluated and compared together in 3 aspects. In the first case, the running time of these methods, in segmenting the land-cover to 2, 3, 4, 5 and 6 segments, are measured and reported in Table 2. It is observed that because of the random nature of methods the ranking of them is changing by change the number of segments. But it can be concluded all methods have same performance, however the proposed method has rapid convergence in the low number of segments.

In the next case study, resulted optimum number of clusters by methods is examined in Table 3 by minimizing the XB cluster validity index. Each method obtain only one number to optimum number of segments, independent from the number of input images.

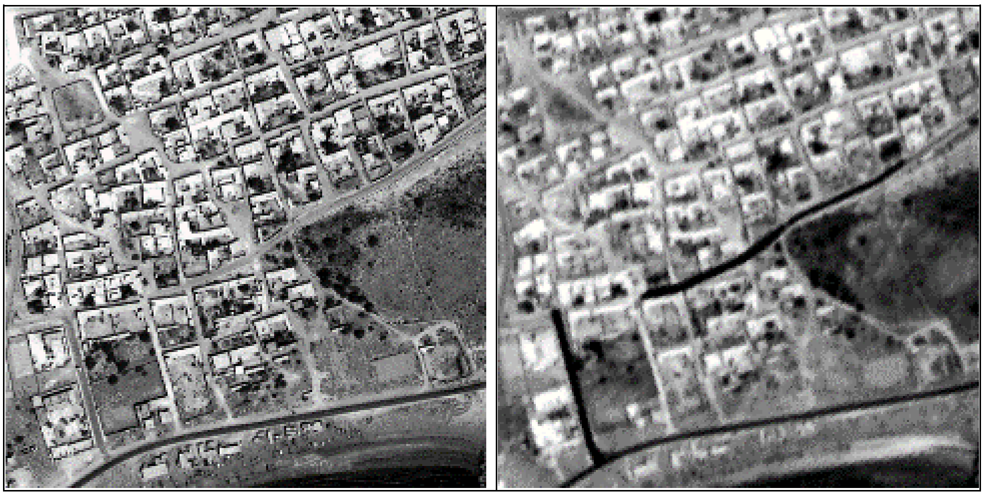
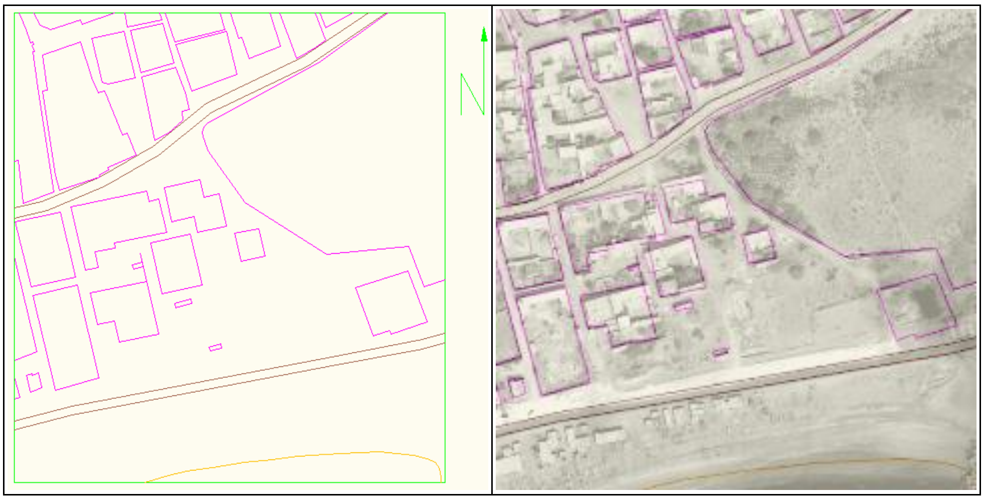


Fig. 8. The second land-cover images (left: A and right: B).



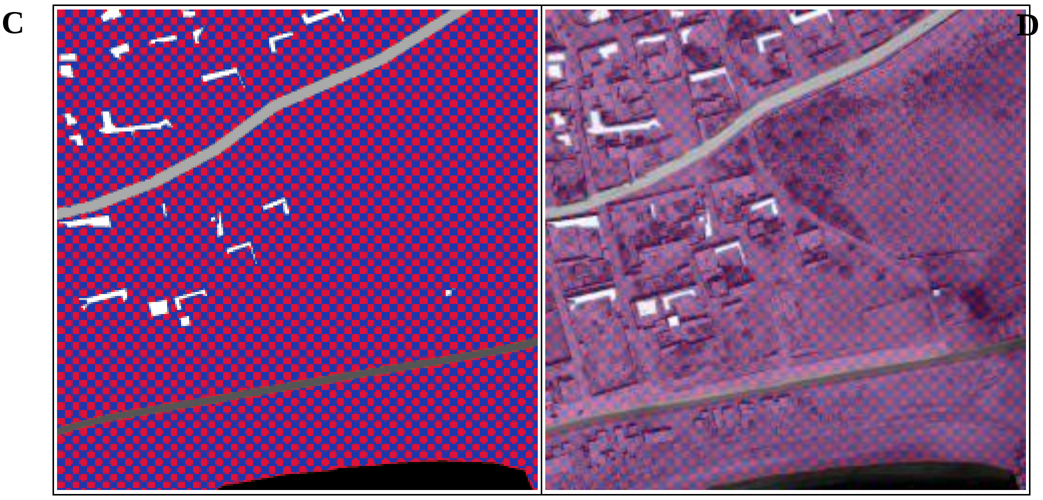


Fig. 9. A: The planimetric map of second land-cover (Figure 8); B: The planimetric map is situated on the Figure 8-A image; C: 4 classes (sea with black DN, main road with new and dark asphalt, sidetrack with old and bright asphalt and building's roof with white DN) are selected and labeled to 1, 2, 3 and 4 respectively on Figure 8-A image, other pixels are Checkered; D: Figure 9-C image is situated on the Figure 8-A image.

In the last study case, the accuracy assessment of these 4 methods and some conventional and new satellite images segmentation methods (Combination HMRF and Expectation Maximization [29], Kernel Level Set Segmentation [45], Multilevel thresholding based on Electro-magnetism Optimization [46] and Harmony Search Optimization [47], Gaussian mixture model [48] and Expectation maximization [49]) will be examined in the segmenting of land-cover to 4 segments. For this purpose, validity indices overall accuracy (OA) and Kappa (Ka) are utilized. Simulation results of this case are reported in Table 4.

It is observed, since the image Figure 8-A is influenced by shadow, furthermore, the slave roads are asphalted partially in the image Figure 8-B, the resulted segmentation by min fusion operator (CFSM-1) has the worst performance. Resulted segmentation by mean fusion operator (CFSM-2) is influenced by the mentioned error cautions (for CFSM-1 too), however less than CFSM-1. Although since the max fusion operator do not influenced by the mentioned error cautions, has the best performance. This arrangement could be inverse, if the error cautions would increase the DN of pixels, e.g. placing the region under the white cloudy dots. Against, the proposed method (PM) always has the acceptable performance, independent from the nature of error cautions. The performance of the other segmentation methods (Combination HMRF and Expectation Maximization [29], Kernel Level Set Segmentation [45], Multilevel thresholding based on Electro-magnetism Optimization [46] and Harmony Search Optimization [47], Gaussian mixture model [48] and Expectation maximization [49]) is similar to CFSM-2, because of averaging performance when they are applied on all scene images. Finally the segmented images by using the proposed and 3 classical fusion based methods are illustrated in Figure 10.

As the conclusions of simulation results, growing the RS satellites in last years, has been increased extremely the available RS data from environment. Nowadays, automatic and body-free processing and exploitation of this massive amount of data is a challenging and interesting subject. On the other hand, the images with nearby DNs to OPIS outcome good results, in the conventional segmentation methods, where they would apply only on the one single image. But there is no general method to get the image with nearby DNs to OPIS. In some regions, there is no OPIS or any knowledge to choice the image with nearest DNs to reality. Therefore, this paper proposes a new method to automatically utilizing of massive available RS data from environment and getting the good response in segmentation. However this good response will differ with obtained response by OPIS, but it is acceptable. This paper is the first work to this purpose. In the future by extending the used fuzzy numbers, metrics and etc. it is going to get the better responses, in the case of using multiple RS images.

6. Conclusions and Future Works

Land-cover segmentation using remotely sensed images is a challenging research topic. This paper proposed a method to fuse multiple panchromatic images in one fuzzy image. Finally, by applying the FCM on the resulted fuzzy image, a single segmented image was obtained for each land-cover. Some reported comparisons and mathematical analysis showed the better performance of the proposed method, in compare to the classical segmentation methods and conventional fusion methods. The performance of the proposed method also was studied when applied on two various scenes (different land-cover type). Simulation results showed the novelty and acceptable performance of the proposed method in noise-free segmentation and run time terms. Since there is a direct relation between the performance of the proposed method and increasing the number of images for each land-cover type, in the future by obtaining more images for these land-covers, it will be tried to increase the robustness of the proposed method versus noises. Furthermore, the simplest state of non-crisp (symbolic interval) numbers and also its simplest correspond metric were used in this paper. By improving them in the future works it will be tried to grow up the performance and efficiency.

Table 2. Running time of proposed and 3 classical fusion based segmentation methods

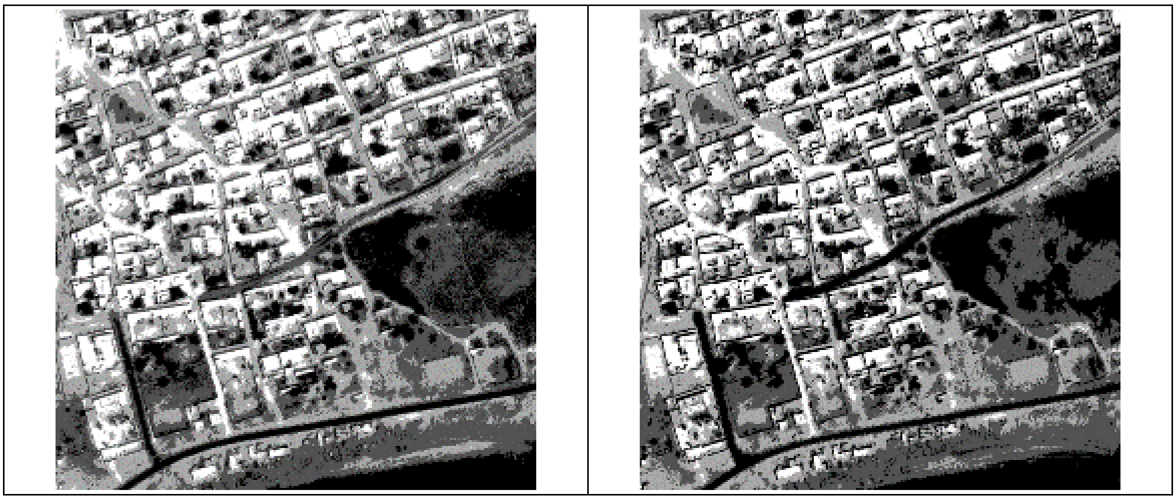
Num. of Seg. Method	2		3		4		5		6		average	
	Rank	Time	Rank	Time	Rank	Time	Rank	Time	Rank	Time	Rank	Time
PM	1	2.756	1	6.954	1	13.214	4	25.072	4	30.43	2	15.69
CFSM-1	2	5.01	2	7.4	4	16.62	2	21.78	2	26.13	1	15.39
CFSM-2	4	5.78	3	9.55	3	16.57	1	21.76	3	26.15	3	15.96
CFSM-3	3	5.64	4	11.6	2	15.65	3	21.78	1	26.05	4	16.14

Table 3. Resulted optimum number of clusters by proposed and 3 classical fusion based segmentation methods by minimizing the XB cluster validity index.

Num. of Seg. Method	2	3	4	5	6	Optimum
	XB	XB	XB	XB	XB	Arg min XB
PM	0.126241	0.141	0.16848	0.165728	0.144674	2
CFSM-1	0.08256	0.071798	0.074929	0.068398	0.07151	5
CFSM-2	0.091861	0.08136	0.078047	0.072338	0.075637	5
CFSM-3	0.097709	0.074982	0.07521	0.075045	0.078609	3

Table 4. Resulted overall accuracy (OA) and Kappa (Ka) indices in the accuracy assessment of proposed, 3 classical fusion based segmentation methods and 6 conventional and new satellite images segmentation methods.

Index Method	%OA		%Ka	
	Rank	Value	Rank	Value
PM	2	57.35	2	41.06
CFSM-1	9	56.33	10	38.25
CFSM-2	3	57.16	5	40.38
CFSM-3	1	68.27	1	56.07
[29]	8	56.53	7	39.51
[45]	4	57.15	4	40.67
[46]	10	55.78	9	38.36
[47]	5	57.11	3	40.84
[48]	6	56.67	6	39.57
[49]	7	56.58	8	39.49



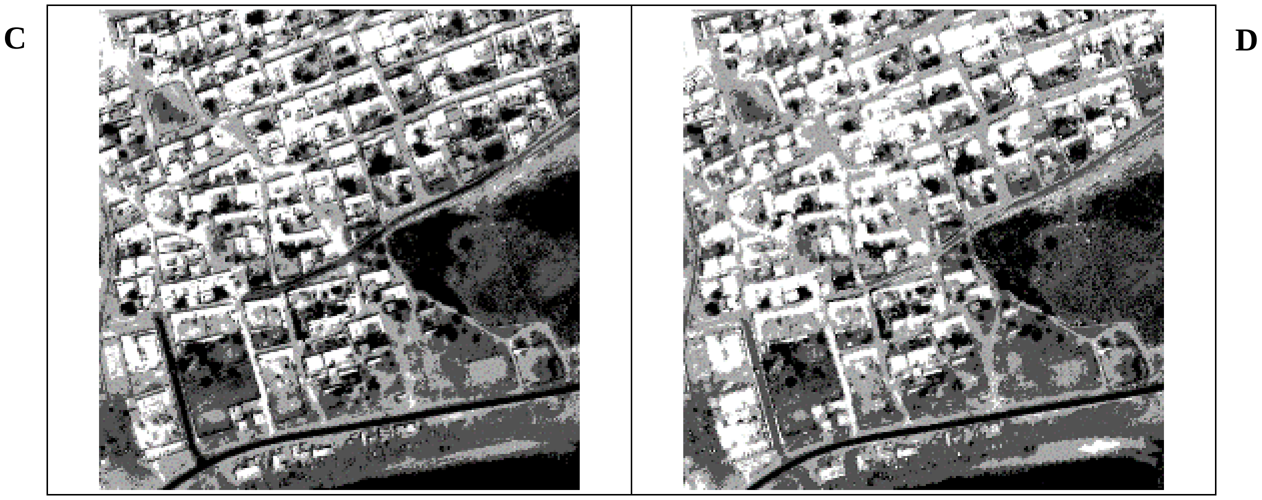


Fig. 10. The resulted images in segmenting of second land-cover by using proposed (A) and 3 classical fusion based segmentation methods (B: CFSM-1, C: CFSM-2, D: CFSM-3).

References

[1] S. Saha and S. Bandyopadhyay, "Application of a New Symmetry-Based Cluster Validity Index for Satellite Image Segmentation," IEEE Geoscience and Remote Sensing Letters, vol. 5, pp. 166-170, 2008.

[2] S. Mitra, M. Dickens, and S. Pemmaraju, "Adaptive Clustering for Segmentation of Multi-sensor Images," 1998.

[3] M. Z. Ahmed REKIK, Ahmed Ben Hamida, Mohammed Benjelloun, "Review of satellite image segmentation for an optimal fusion system based on the edge and region approaches," International Journal of Computer Science and Network Security, vol. 7, pp. 242-250, 2007.

[4] I. Saha, U. Maulik, S. Bandyopadhyay, and D. Plewczynski, "SVMeFC: SVM Ensemble Fuzzy Clustering for Satellite Image Segmentation," Geoscience and Remote Sensing Letters, IEEE, vol. 9, pp. 52-55, 2012.

[5] G. Wang, G. Z. Gertner, S. Fang, and A. B. Anderson, "A methodology for spatial uncertainty analysis of remote sensing and GIS products," Photogrammetric Engineering & Remote Sensing, vol. 71, pp. 1423-1432, 2005.

[6] C. T. Hunsaker, M. F. Goodchild, M. A. Friedl, and T. J. Case, Spatial uncertainty in ecology: implications for remote sensing and GIS applications: Springer Science & Business Media, 2013.

[7] H. T. Mowrer and R. G. Congalton, Quantifying spatial uncertainty in natural resources: theory and applications for GIS and Remote Sensing: CRC Press, 2003.

[8] T. Cheng, "Fuzzy objects: their changes and uncertainties," Photogrammetric Engineering and Remote Sensing, vol. 68, pp. 41-50, 2002.

[9] F. Wang and G. B. Hall, "Fuzzy representation of geographical boundaries in GIS," International Journal of Geographical Information Systems, vol. 10, pp. 573-590, 1996.

[10] G. Foody, "Fuzzy modelling of vegetation from remotely sensed imagery," Ecological modelling, vol. 85, pp. 3-12, 1996.

[11] X. Zhao, A. Stein, and X. Chen, "Application of random sets to model uncertainties of natural entities extracted from remote sensing images," Stochastic Environmental Research and Risk Assessment, vol. 24, pp. 713-723, 2010.

[12] P. Fisher, T. Cheng, and J. Wood, "Higher order vagueness in geographical information: Empirical geographical population of type n fuzzy sets," Geoinformatica, vol. 11, pp. 311-330, 2007.

[13] T. Cheng and M. Molenaar, "Objects with fuzzy spatial extent," Photogrammetric Engineering and Remote Sensing, vol. 65, pp. 797-802, 1999.

[14] G. Rees, Physical principles of remote sensing vol. 1: Cambridge Univ Pr, 2001.

[15] E. Yasunori, T. Isao, H. Yukihiro, and M. Sadaaki, "Kernelized fuzzy c-means clustering for uncertain data using quadratic penalty-vector regularization with explicit mappings," in Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, 2011, pp. 804-809.

[16] B. Kao, S. D. Lee, D. W. Cheung, W.-S. Ho, and K. Chan, "Clustering uncertain data using voronoi diagrams," in Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, 2008, pp. 333-342.

[17] G. B. Heuvelink, J. D. Brown, and E. Van Loon, "A probabilistic framework for representing and simulating uncertain environmental variables," International Journal of Geographical Information Science, vol. 21, pp. 497-513, 2007.

[18] X. Yu, H. He, D. Hu, and W. Zhou, "Land cover classification of remote sensing imagery based on interval-valued data fuzzy c-means algorithm," Science China Earth Sciences, vol. 57, pp. 1306-1313, 2014/06/01 2014.

[19] A. Stein, N. Hamm, and Q. Ye, "Handling uncertainties in image mining for remote sensing studies," International journal of remote sensing, vol. 30, pp. 5365-5382, 2009.

[20] Y. El-Sonbaty and M. A. Ismail, "Fuzzy clustering for symbolic data," IEEE Transaction on Fuzzy Systems, vol. 6, pp. 195–204, 1998.

[21] S. Lee and M. Crawford, "Unsupervised classification for multi-sensor data in remote sensing using Markov random field and maximum entropy method," in IEEE 1999 International Geoscience and Remote Sensing Symposium, 1999. IGARSS'99 Proceedings., 1999, pp. 1200-1202.

[22] S. Lee, A. Suh, and M. Jung, "Multi-sensor data classification in remote sensing using MRF regional growing algorithm," in IEEE 2001 International Geoscience and Remote Sensing Symposium, 2001. IGARSS'01., 2001, pp. 2884-2886.

[23] S. Lee and M. Crawford, "Multi-channel/multi-sensor image classification using hierarchical clustering and fuzzy classification," in IEEE 2000 International Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000., 2000, pp. 957-959.

[24] S. Nazarko, "Evaluation of data fusion methods using Kalman filtering and transferable belief model," Master’'s thesis, University of Jyväskylä, 2002.

[25] X. Dai and S. Khorram, "Data fusion using artificial neural networks: a case study on multitemporal change analysis," Computers, Environment and Urban Systems, vol. 23, pp. 19-31, 1999.

[26] H. Ghassemian, "Multisensor Image Fusion by Inverse Subband Coding," Proceeding of ISPRS-2000, CD, vol. 3.

[27] M. Hasanzadeh and S. Kasaei, "A Multispectral Image Segmentation Method Using Size-Weighted Fuzzy Clustering and Membership Connectedness," IEEE Geoscience and Remote Sensing Letters, vol. 7, 2010.

[28] A. A. Naeini, S. Niazmardi, S. R. Namin, F. Samadzadegan, and S. Homayouni, "A Comparison Study Between Two Hyperspectral Clustering Methods: KFCM and PSO-FCM," in Computational Intelligence and Decision Making, ed: Springer, 2013, pp. 23-33.

[29] A. Bendjebbour, L. Fouque, V. Samson, and W. Pieczynski, "Multisensor Image Segmentation Using Dempster–Shafer Fusion in Markov Fields Context," IEEE Transaction on Geoscience and Remote Sensing, vol. 38, pp. 1789-1798, 2001.

[30] B. Benmiloud and W. Pieczynski, "Estimation des parameters dans les chaines de Markov cachees et segmentation d'images," Traitement du signal, vol. 12, pp. 433-454, 1995.

[31] N. Giordana and W. Pieczynski, "Estimation of generalized multisensor hidden Markov chains and unsupervised image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 465-475, 1997.

[32] A. Bendjebbour and W. Pieczynski, "Multisensor Evidential Hidden Markov Fields and Image Segmentation," presented at the Second IEEE Interantional Conference on Intelligent Processing Systems (ICIPC'98), Australia, 1998.

[33] N. Giordana and W. Pieczynski, "Unsupervised segmentation of multisensor images using generalized hidden Markov chains," in International Conference on Image Processing, 1996. Proceedings., 1996, pp. 987-990 vol.3.

[34] M. S. Balch, "Methods for Rigorous Uncertainty Quantification with Application to a Mars Atmosphere Model," Virginia Polytechnic Institute and State University, 2010.

[35] M. Hadi, K. Morteza, and S. Y. Hadi, "Vector fuzzy C-means," Journal of Intelligent and Fuzzy Systems, vol. 24, pp. 363-381, 2013.

[36] K. L. Wu and M. S. Yang, "Alternative c-means clustering algorithms," Pattern Recognition, vol. 35, pp. 2267-2278, 2002.

[37] C. C. Chuang, J. T. Jeng, and C. W. Li, " Fuzzy C-Means Clustering Algorithm with Unknown Number of Clusters for Symbolic Interval Data," presented at the SICE Annual Conference, 2008.

[38] S. Saha and S. Bandyopadhyay, "Application of a Multiseed-Based Clustering Technique for Automatic Satellite Image Segmentation," IEEE Geoscience and Remote Sensing Letters, vol. 7, pp. 306-308, 2010.

[39] S. Das and S. Sil, "Kernel-induced fuzzy clustering of image pixels with an improved differential evolution algorithm," Information Sciences, vol. 180, pp. 1237–1256, 2010.

[40] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," IEEE Transactions on pattern analysis and machine intelligence, vol. 13, pp. 841-847, 1991.

[41] S.-B. Cho and S.-H. Yoo, "Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data," Pattern recognition, vol. 39, pp. 2405-2414, 2006.

[42] P. D'Urso and P. Giordani, "A weighted fuzzy c-means clustering model for fuzzy data," Computational Statistics & Data Analysisvol, vol. 50, pp. 1496-1523, 2006.

[43] F. A. T. D. Carvalho, "Fuzzy clustering algorithms for symbolic interval data based on adaptive and non-adaptive Euclidean distances," in Proceedings of the Ninth Brazilian Symposium on Neural Networks (SBRN'06), 2006, pp. 60-65.

[44] F. A. T. D. Carvalho, "Fuzzy c-means clustering methods for symbolic interval data," Pattern Recognition Letters, vol. 28, pp. 423–437, 2007.

[45] M. Ben Salah, A. Mitiche, and I. Ben Ayed, "Effective level set image segmentation with a kernel induced data term," Image Processing, IEEE Transactions on, vol. 19, pp. 220-232, 2010.

[46] D. Oliva, E. Cuevas, G. Pajares, D. Zaldivar, and V. Osuna, "A Multilevel Thresholding algorithm using electromagnetism optimization," Neurocomputing, vol. 139, pp. 357-381, 2014.

[47] D. Oliva, E. Cuevas, G. Pajares, D. Zaldivar, and M. Perez-Cisneros, "Multilevel thresholding segmentation based on harmony search optimization," Journal of Applied Mathematics, vol. 2013, 2013.

[48] H. Greenspan, A. Ruf, and J. Goldberger, "Constrained Gaussian mixture model framework for automatic segmentation of MR brain images," Medical Imaging, IEEE Transactions on, vol. 25, pp. 1233-1245, 2006.

[49] M. L. Comer and E. J. Delp, "The EM/MPM algorithm for segmentation of textured images: Analysis and further experimental results," Image Processing, IEEE Transactions on, vol. 9, pp. 1731-1744, 2000.

Hadi Mahdipour Hossein-Abad was born in Shirvan, Iran, in 1984. He received the B.S. degree in electrical engineering (Communication trend) from Ferdowsi University of Mashhad, Iran, in 2005. Then he received the M.S. degree in electrical engineering (Communication trends) from Shahid Bahonar University of kerman, Iran, in 2007. Finally he received the PHD degree in electrical engineering (Communication-sysem trend) from Ferdowsi University of Mashhad, Iran, in 20۱۵. He is currently Assisant Professor at Faculty of Marine Engineering, Khorramshahr University of Marine Science and Technology, Iran, from 20۱۵. His research interests include pattern recognition, image and video processing and telecommunication. Email: mahdipour@kmsu.ac.ir.

Morteza Khademi was born in Iran, in 1958. He received the B.Sc. and M.S. degrees from Isfahan University of Technology, Isfahan, Iran, in 1985 and 1987, respectively and the Ph.D. degree from the Wollongong University, Australia, on video communications in 1995, all in Electrical Engineering. He joined Ferdowsi University of Mashhad, Iran in 1987. He is currently Professor at the Department of Electrical Engineering, Ferdowsi University of Mashhad, Iran. His research interests include image and video processing and communication and biomedical signal processing. Email: khademi@um.ac.ir.

Hadi Sadoghi Yazdi received the B.S. degree in electrical engineering from Ferdowsi University of Mashhad, Iran in 1994, and then he received to the M.S. and PhD degrees in electrical engineering from Tarbiat Modarres University of Tehran, Iran, in 1996 and 2005 respectively. He works in Computer Department as a professor at Ferdowsi University of Mashhad. His research interests include adaptive filtering, image and video processing, and optimization in signal processing. Email: h-sadoghi@um.ac.ir.

Preserving Data Clustering with Expectation Maximization Algorithm

Leila Jafar Tafreshi*
Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran
leila.tafreshi66@yahoo.com
Farzin Yaghmaee
Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran
f_yaghmaee@semnan.ac.ir

Received: 30/Jun/2015 Revised: 03/Aug/2016 Accepted: 13/Aug/2016

Abstract

Data mining and knowledge discovery are important technologies for business and research. Despite their benefits in various areas such as marketing, business and medical analysis, the use of data mining techniques can also result in new threats to privacy and information security. Therefore, a new class of data mining methods called privacy preserving data mining (PPDM) has been developed. The aim of researches in this field is to develop techniques those could be applied to databases without violating the privacy of individuals. In this work we introduce a new approach to preserve sensitive information in databases with both numerical and categorical attributes using fuzzy logic. We map a database into a new one that conceals private information while preserving mining benefits. In our proposed method, we use fuzzy membership functions (MFs) such as Gaussian, P-shaped, Sigmoid, S-shaped and Z-shaped for private data. Then we cluster modified datasets by Expectation Maximization (EM) algorithm. Our experimental results show that using fuzzy logic for preserving data privacy guarantees valid data clustering results while protecting sensitive information. The accuracy of the clustering algorithm using fuzzy data is approximately equivalent to original data and is better than the state of the art methods in this field.

Keywords: Privacy Preserving; Clustering; Data Mining; Expectation Maximization Algorithm.

1. Introduction

Huge volumes of individual’s information are frequently collected and analyzed by various applications such as shopping habits, criminal records, medicinal history and credit records. On the other hand, data has an important role in decision making for business organizations and governments. Therefore, analyzing such data may threaten individual's privacy. Also, companies that generate a huge burden of data often need to transmit these data to third parties for their studies. As data usually contains sensitive information about people and corporations, their releasing to third parties requires mechanisms to make sure that data privacy is preserved [1].

For example, a bank may release credit records of individuals for statistical purpose or a hospital may release patient’s diagnosis records. Both of these applications need the individual's data to be private while data mining process.

However, there may exist other attributes that can be used, in combination with an external database, to recover the personal identities.

Privacy preserving data mining entails two notions:

I. Extracting or mining knowledge from large amounts of data.
II. Performing data mining in such a way that data privacy is not compromised.

For these purposes, we should preserve the privacy of data or the knowledge discovered from mining results [2].

The extracted knowledge form data is generally expressed in the form of clusters, decision trees or association rules which allows one to mine the information. In this work we focus on clustering methods.

Clustering is known as identification of similar objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution patterns and correlations between data attributes.

In privacy preserving data while clustering, the main goal is to find the clusters of data without revealing the content of data elements themselves.[3] It is important to use a kind of modification technique in order to achieve higher accuracy of data mining results.

Most of works in privacy preserving clustering are developed on k-means algorithm by applying the model of secure multi-party computation on different distributions [4].

Data modification techniques for PPDM can be classified into two principle groups: perturbation-based and anonymization-based techniques according to how the protection of privacy.

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. It even assumes that sensitive data should be retained for analysis.

* Corresponding Author

Perturbation of data is a very easy and effective method for protecting the sensitive information of the data from unauthorized users or hackers. There are two types of data perturbation for protecting data namely:

- Value-based Perturbation: the purpose is to preserve statistical characteristics and columns distribution. This approach perturbs data by adding noise, or other randomized processes.
- Multi-Dimensional Perturbation: aims to hold Multi-Dimensional information. This approach includes random projection or random rotation techniques. Comparing to other multi-dimensional data perturbation methods, these perturbations exhibit unique properties for privacy preserving data classification and clustering. [6]

All of these methods introduce a bit of complexity in their algorithms. Our main goal is to introduce a simple and optimum solution for privacy preserving data mining problem using fuzzy logic.

In this research we compare our clustering results with some other methods such as random projection, random rotation and noise addition.

In this work we have applied the idea of using fuzzy logic to preserve the individual's information while revealing details in public. In this regard, we used one of the fuzzy membership functions described in table (1) to transform original data. Then transformed data were clustered by EM algorithm to evaluate our method.

The rest of the paper is organized as follows: in section 2 we describe related works in privacy preserving data mining. Section 3 presents state of the art works which we used them to compare the experimental results. Section 4 explains the proposed method based on fuzzy logic. Section 5 and 6 is dedicated to experimental results and conclusion respectively.

Table 1. Fuzzy membership functions

<table>
<tr><th colspan="2"><i>P-shaped MF</i></th><th colspan="2"><i>Z–shaped MF</i></th><th colspan="2"><i>S–shaped MF</i></th></tr>
<tr>
<td colspan="2">F (x؛a,b,c,d) =
$0, x \le a$
$a \le x \le \frac{a+b}{2} \to 2\left(\frac{x-b}{b-a}\right)^2$
$\frac{a+b}{2} \le x \le b \to 1-2\left(\frac{x-b}{b-a}\right)^2$
$b \le x \le c \to 1$
$c \le x \le \frac{c+d}{2} \to 1-2\left(\frac{x-c}{d-c}\right)^2$
$\frac{c+d}{2} \le x \le d \to 2\left(\frac{x-d}{d-c}\right)^2$
$x \ge d \to 0$
The parameters a and d locate the "feet" of the curve, while b and c locate its "shoulders.</td>
<td colspan="2">F (x؛a,b) =
$x \le a \to 1$
$a \le x \le \frac{a+b}{2} \to 1-2\left(\frac{x-a}{b-a}\right)^2$
$\frac{a+b}{2} \le x \le b \to 2\left(\frac{x-b}{b-a}\right)^2$
$x \ge b \to 0$
The parameters a and b locate the extremes of the sloped portion of the curve.</td>
<td colspan="2">F (x؛a,b) =
$x \le a \to 0$
$a \le x \le \frac{a+b}{2} \to 2\left(\frac{x-b}{b-a}\right)^2$
$\frac{a+b}{2} \le x \le b \to 1-2\left(\frac{x-b}{b-a}\right)^2$
$x \ge b \to 1$
The parameters a and b locate the extremes of the sloped portion of the curve.</td>
</tr>
<tr><th colspan="3"><i>Gaussian MF</i></th><th colspan="3"><i>Sigmoid MF</i></th></tr>
<tr>
<td colspan="3">$gaussian(x; c, \sigma) = e^{-\frac{1(\frac{x-c}{\sigma})^2}{2}}$
Parameter c represents the MFs center and σ determines the MFs width.</td>
<td colspan="3">$sig(x; a, c) = \frac{1}{1+\exp[-a(x-c)]}$
Parameter a, controls the slope at the crossover point b.</td>
</tr>
</table>

2. Related Works

The term, privacy-preserving data mining, introduced by Agrawal and Srikant in 2000 [7]. The initial idea of PPDM is extending traditional data mining techniques to work with the modified data those mask sensitive information. The key issues is how to modify data and how to recover data mining results from the modified data. The solutions are often tightly coupled with the data mining algorithms.

One of such techniques is Rotation-Based Transformation (RBT) [8]. A novel spatial data transformation method for Privacy Preserving Clustering. This method is designed to protect the underlying attribute values subjected to clustering without jeopardizing the similarity between data objects under analysis. Releasing a database transformed by RBT, a database owner meets privacy requirements and guarantees valid clustering results. The data is shared after the transformation to preserve privacy without normalization. Researches show that having previous knowledge, the random rotation perturbation may become involved in privacy violations against different attacks including Independent Component Analysis (ICA), attack to rotation center and distance-inference attack.

Another work is Random Projection-Based, which is a dimension reduction technique, introduced by Kun Liu, Hillol Kargupta and Jessica Ryan in 2006 [9]. This work uses random projection matrices which is a tool for PPDM. It proves that, after perturbation, the distance-related statistical properties of the original data is well maintained without divulging the dimensionality and the exact data values. The experimental results demonstrate that this technique can be successfully applied to different kinds of data mining tasks such as inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier Detection and linear classification.

However, this technique can hardly preserve the distance and inner product during the modification in comparison with geometric and random rotation techniques. It has been also clarified that having previous knowledge about Random Projection-Based perturbation technique may be caught into privacy breach against the attacks. Our purposed technique does not loose data which is its benefit versus random projection.

Another work for privacy preserving clustering is double reflecting data perturbation and rotation data perturbation which is proposed in [10].

Kadampur and Somayajulu presented a method of privacy preserving clustering by cluster bulging [11]. In this method, the original values of individual objects are not revealed and the privacy of individual objects is preserved; but the perturbed dataset is still relevant for cluster analysis.

Yifeng and Harbin combined the random response technology and the geometric data transformation method in 2009 which is called random response method of geometric transformation [12]. It can protect the privacy of numerical data. Theoretical analysis and experimental results show that the algorithm improves privacy protection than the previous algorithms.

In [13] a family of geometric data transformation methods (GDTMs) which ensure that the mining process up to a certain degree of security is introduced. Their method is designed to address the privacy preservation in clustering analysis, This method distort only confidential numerical attributes to meet privacy requirements, while preserving general features for clustering analysis.

Shibnath Mukherjee, Zhiyuan Chen and Aryya Gangopadhyay proposed an integrated Dimension Reduction-based approach for data reduction and privacy for distance-based mining algorithms using Fourier-related transform [14]. Experimental results demonstrate that the proposed approach leads to much better mining quality than the existing random perturbation and random projection approaches given the same degree of privacy in both centralized and distributed cases.

Shalini Lamba present a potential approach to preserve the individual's details by transforming the original data into fuzzy data [15]. They have used only numerical data for their experimentation purpose. The main goal of their technique is reducing the run time of preserving data privacy while clustering.

3. State of the Art Methods

In this section we explain the three famous methods in this field which we used them to compare the experimental results of proposed method.

3.1 Expectation Maximization

Expectation Maximization (EM) is a well-established clustering algorithm in statistics community. EM is a distance-based algorithm which assumes that the dataset can be modeled as a linear combination of multivariate normal distributions and the algorithm finds the distribution parameters that maximizes a model quality measure, called log likelihood.

EM is linear in database size, robust to noisy data, can handle high dimensionality and has a very good quality while using huge datasets.

This algorithm assumes Apriori that tries to fit the data into 'n' Gaussian channel by expecting the classes of all data point and finding the maximum likelihood of Gaussian centers. Algorithmic steps for EM is as follows:

Let $x = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points, $v = \{\mu_1, \mu_2, \mu_3, \dots, \mu_c\}$ be the set of means of Gaussian.

$p = \{p_1, p_2, p_3, \dots, p_c\}$ is the set of probability of occurrence of each Gaussian.

1. On the i$^{t'h}$ iteration initialize:

$$\lambda_t = \{\mu_1(t), \dots \mu_c(t), \Sigma_1(t), \Sigma_2(t), .. \Sigma_c(t), p_1(t), \dots p_c(t)\} \tag{1}$$

2. Compute the "expected" classes of all data points for each class using:

$$p(w_t | x_k, \lambda_t) = \frac{p(x_k | w_t, \lambda_t) p(w_t | \lambda_t)}{p(x_k | \lambda_t)} = \frac{p(x_k | w_t, \mu_t(t), \Sigma_t(t))\, p_t(t)}{\sum_{j=1}^{c} p(x_k | w_j, \mu_j(t), \Sigma_j(t))\, p_j(t)} \tag{2}$$

3. Compute the maximum likelihood (μ) given our data class membership distribution using:

$$\mu_t(t+1) = \frac{\sum_k p(w_t | x_k, \lambda_t) x_k}{\sum_k p(w_t | x_k, \lambda_t)} \tag{3}$$

$$p_t(t+1) = \frac{\sum_k p(w_t | x_k, \lambda_t)}{R} \tag{4}$$

Where *'R'* is the number of data points. Repeat the steps 2 and 3 while algorithm converges.

3.2 Random Projection

Random projection [9] refers to a technique of projecting a set of data points from high-dimensional space to a randomly chosen lower-dimensional subspace.

If the matrix $X\ m \times n (or\ Y\ m \times n)$ indicates original dataset, $R_{n \times k} (k < n) (or\ R'_{k \times m} (k < m))$ is a random matrix such that each entry $r_{i\times j}$ of $R(or\ R')$ is independent and identically chosen from some unknown distribution with mean zero and variance σ_r^2, the Column-wise Projection *G(X)* and Row-wise Projection *G(Y)* will be defined as below:

$$G(X)=\frac{1}{\sqrt{k\sigma_r}}XR, G(Y)=\frac{1}{\sqrt{k\sigma_r}}R'Y \tag{5}$$

The key idea of random projection arises from the Johnson-Linden Strauss Lemma. According to this lemma, it is possible to maintain distance-related statistical properties simultaneously with dimension reduction for a dataset. Therefore, this perturbation technique can be used for different data mining tasks like including inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier detection, linear classification, etc. [16].

This method reduces the dimensionality of data by projecting it onto a lower dimensional subspace using a random matrix with columns of unit length [17].

3.3 Random Rotation

This category includes all orthonormal perturbations. $R_{d\times d}$ Represent the rotation matrix. Geometric rotation of the data X is as a function $f(X), f(X) = RX$. Transformation will not change the label of data tuples. $R_{d\times d}$ Have the following properties.

Let R^T represent the transpose of the matrix R, r_{ij} represent the (i, j) element of R, and I be the identity matrix. Both the rows and the columns of R , are orthonormal i.e., for any column $j, \sum_{i=1}^{d} r\ ij^2 = 1$ and for any two columns j and k, $\sum_{i=1}^{d} r_{ij}r_{ik} = 0$. The similar property is held for rows. The definition infers that $R^TR = RR^T = I$. It also implies that by changing the order of the rows or columns of rotation matrix, the resulting matrix is still a rotation matrix. A random rotation matrix can be efficiently generated following the “Haar” distribution. For example the Eq. (6) is a rotation matrix.

$$R = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \tag{6}$$

The goals of rotation based data perturbation are: preserving the accuracy of classifiers and preserving the privacy of data.

A key feature of rotation transformation is preserving length and the Euclidean distance between any pair of points x and y, the inner product is also invariant to rotation.

3.4 Random Noise Addition Technique

This technique is described in as follows: Consider ^{n} original data $X_1, X_2, \dots, X_N$, where X_i are variables following the same independent and identical distribution. The distribution function of X_i is denoted as F_x, n random variables $Y_1, Y_2, \dots, Y_N$ are generated to hide the real values of X_i by perturbation. Disturbed data will be generated as:

$$w_i = X_i + Y_i \text{ Where } i = 1, \dots, n \tag{7}$$

It is also assumed that the added noise variance is large enough to let an accurate estimation of main data values take place. Then, according to the perturbed dataset $w_1, \dots, w_n$ known distributional function FY and using a reconstruction procedure based on Bayes rule, the density function $f'x$ will be estimated by Equation:

$$f'_X(a)=\frac{1}{n}\sum_{i=1}^{n}\frac{\int_{-\infty}^{a} f_Y(w_i-a)f_X(a)}{\int_{-\infty}^{-\infty} f_Y(w_i-z)f_X(z)d_z} \tag{8}$$

4. Proposed Method

In this work we propose a new approach to preserve sensitive data in general databases which have both numerical and categorical attributes with clustering by Expectation Maximization algorithm. First of all, we map categorical private data into numerical data and second the numerical private data are transformed using fuzzy membership functions described in Table (1). Finally, we cluster the transformed data using EM algorithm. For better clarity we describe our process step by step.

- Categorical private attributes are mapped to numerical values. This can be done by a simple method which works on ASCII codes of alphabet characters of each field.
- Private attribute’s values are transformed using fuzzy membership functions and fuzzy data are sent back to the user.
- The received data are grouped into different clusters using EM.

As mentioned above, in second phase we transform private attribute’s values to fuzzy values using a fuzzy membership function such as P-shaped, Z-shaped, S-shaped, Gaussian and Sigmoid, which are described in Table (1). Fuzzy logic is an approach to compute based on "degrees of truth" rather than usual "true or false" (0 or 1) Boolean logic. Fuzzy logic has been employed to handle the concept of partial truth, where the truth values may lie between complete truth or complete false.

By fuzzifying private data through a fuzzy membership function, each point in the input space is mapped to a value between 0 and 1. So, no one can find the real values of the private data. In the next section we show that this data transformation do not considerably change the mining results.

In the following we describe the datasets and data mining software which we used.

For our experimental purpose, we have used the Weka software which is a popular software for machine learning applications [19].

We have clustered the following described datasets by Weka software to compare the clustering accuracy of fuzzy data with state of the art techniques. We have used datasets with private attributes which should be preserved in data mining process.

The datasets are Pima Indians, German Credit, Student Evaluation, Census Income and Bank Marketing are taken from the UCI repository. These datasets have different number of instances and attributes. Also, they have various attribute types and different number of private attributes.

Pima Indians dataset, Includes cost data (donated by Peter Turney) from National Institute of Diabetes and Digestive and Kidney Diseases. German credit describes the German people credit information. Student evaluation dataset contains 5820 evaluation scores provided by students from Gazi University in Ankara (Turkey). The Census Income (ADULT) dataset predicts whether income exceeds $50k/y or not. Bank marketing dataset is related to direct marketing campaigns of a Portuguese banking institution. The above mentioned datasets details are described in Table (2).

Table (3) shows an example of transforming a record of Census Income dataset using Gaussian fuzzy membership function. The first row shows a part of the original record, second row shows the mapped record and the third row shows the transformed record using Gaussian fuzzy membership function.

Table 2. Datasets description

Datasets	*Number of Records*	*Number of Attributes*	*Number of Private Attributes*	*Number of Clusters*
Pima Indians	768	8	2	2
German Credit	1000	20	5	2
Student Evaluation	5820	33	4	13
Census Income	11012	14	7	2
Bank Marketing	41188	20	4	2

Table 3. An example of transforming private data using Gaussian mf

	Age	*Work class*	*Education*	*Marital-status*	*Occupation*	*Relationship*
Original data	34	Local-gov	Some-college	Never-married	Protective-serve	Not-in-family
Mapped data	34	1	4	3	10	4
Transformed data with Gaussian mf	0.185	0.167	0.186	0.2	0.609	0.184

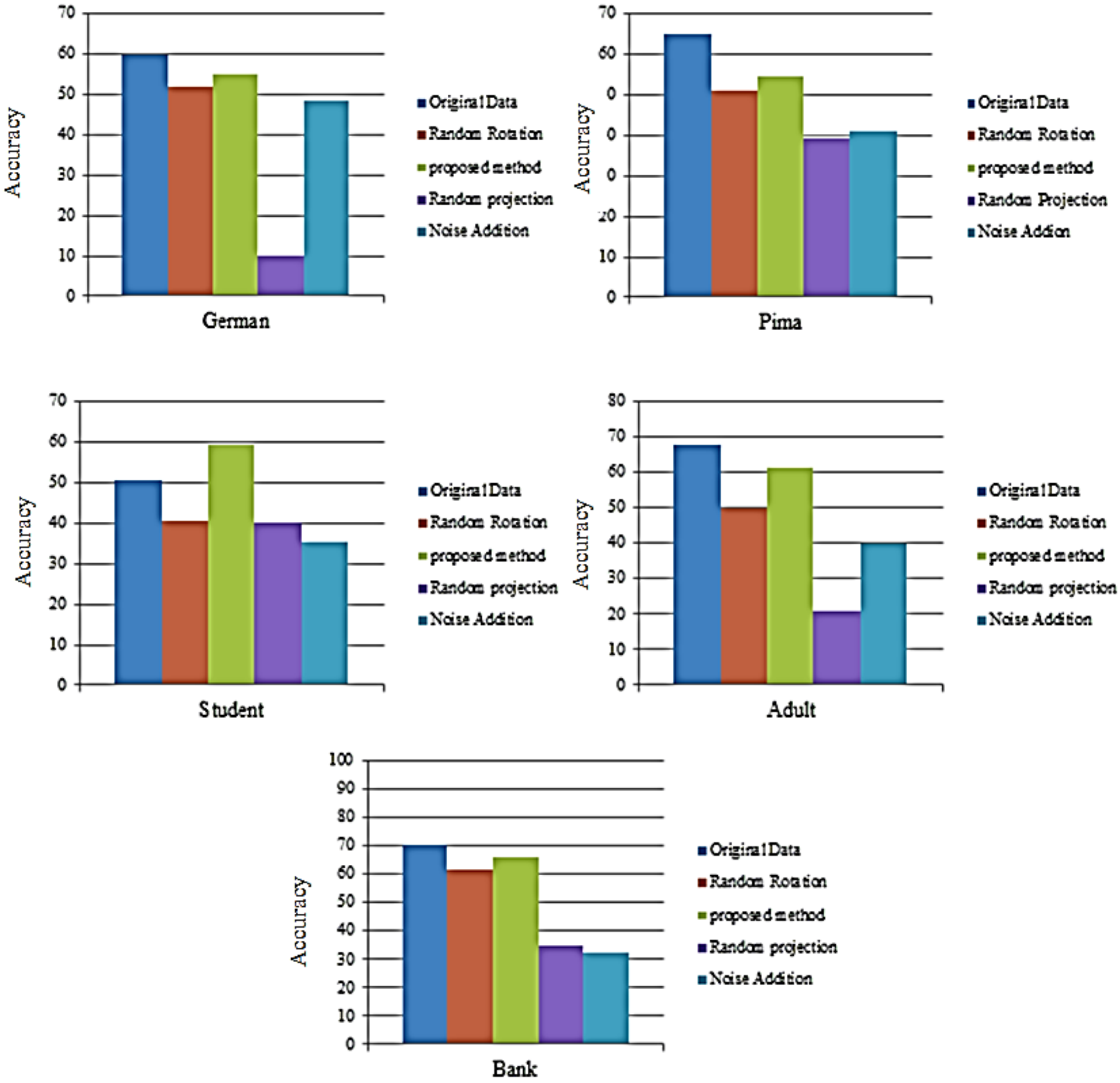


Fig. 1. Comparing the accuracy of EM clustering algorithm while transforming data using different fuzzy memberships

5. Experimental Results

Although, applying different fuzzy membership functions may lead to different clustering accuracy; in this study we used the average result of applying five membership functions described in Table (1) on private data for a fair comparison. As presented in Fig.1, we found that the accuracy of EM algorithm through fuzzy data in all datasets is better than the clustering accuracy of random rotated, random projected and noise added data.

We clustered fuzzified data using EM algorithm. As shown in fig. 1, the clustering accuracy is not changing considerably in compared with those of the original data. This can be due to the fuzzy logic feature which maps each input value to a value between 0 and 1. In addition, using fuzzy logic, data privacy is preserved and no one can guess the original data from the fuzzified data.

By using other methods such as random rotation, random projection and noise addition, clustering accuracy decreases and data privacy is not preserved appreciably.

For noise addition technique we used the average of 20%, 30%, 50% and 60% random noise addition to private data and for random projection method we used the average decreasing about 70%, 40%, 20% and 10% of the dataset's dimension. In Fig. 1 we compared the clustering accuracy of fuzzy data with these two algorithms and random rotation algorithm as well.

However we analyzed the results of proposed method with different fuzzy membership as described in Table (1). Our experimental results show that the shape of fuzzy membership function used to transform original data has not a meaningful effect on EM clustering accuracy. This means that we can use simple fuzzy membership function to reduce the computation complexity.

6. Conclusions

Releasing data and extracting knowledge without violating individual's privacy are important and complicated problems. Most of the existing methods preserve data privacy via complicated processes with disadvantages such as making essential changes or losing data; so they lose the benefits of mining. In this work we focused primarily on privacy issues in data mining, notably when data are revealed for clustering with Expectation Maximization algorithm.

We transformed private data by using fuzzy membership functions to convert a database into a new one in such a way as to preserve the main features of the original database for mining with Expectation Maximization algorithm.

Using fuzzy logic, the relationship between data is maintained while no losing data is occurred.

The results show that our method improves more than 3 percent in clustering accuracy in comparison with other conventional models, such as random rotation, random projection and noise addition.

Another important benefit of proposed algorithm is that by using fuzzy logic, user can tradeoff between privacy preserving and data mining results. This means that proposed method has better flexibility in real applications which requires different security levels.

Future works can be focuses on different clustering algorithms or using encryption algorithms to improve the security requirements.

References

[1] NIRT, RGPV, and Sajjan Singh Nagar, "A review paper on Privacy-Preserving Data Mining." Scholars Journal of Engineering and Technology (SJET), Vol. 1, No. 3, pp. 117-121, 2013.

[2] Lokesh Patel, Prof. Ravindra Gupta, "A Survey of Perturbation Technique for Privacy-Preserving of Data." International Journal of Emerging Technology and Advanced Engineering, Vol. 3, No. 6, pp. 162-166, 2013.

[3] Jharna Chopra, Sampada Satav, "Privacy preservation techniques in data mining." International Journal of Research in Engineering and Technology (IJRET), Vol. 2, No. 4, pp. 537-541, Year 2013.

[4] Tamanna Kachwala, Dr. L. K. Sharma, "A Literature analysis on Privacy Preserving Data Mining." International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 4, April 2015.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society, Series B, Vol. 39, No. 1, pp. 1–38,1977.

[6] Ronica Raj, Veena Kulkarni, "A Study on Privacy Preserving Data Mining: Techniques, Challenges and Future Prospects." International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015.

[7] R. Agrawal and R. Srikant. "Privacy Preserving DataMining." In Proc. ACM SIGMOD Conference on Management of Data, Dallas, Texas, 2000, pp. 439-450.

[8] S. R. M. Oliveira, and O. R.Zaiyane, "Achieving Privacy Preservation When Sharing Data for Clustering." In Proc. Workshop on Secure Data Management in a Connected World, in conjunction with VLDB, Toronto, Ontario, Canada, 2004, pp. 67–82.

[9] Kun Liu, Hillol Kargupta, Senior Member, IEEE, and Jessica Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining." IEEE transactions on knowledge and data engineering, Vol. 18, No. 1, PP. 92-106, 2006.

[10] Liming Li, Sch. of Manage, Fuzhou Univ, Fuzhou, Qishan Zhang, "A Privacy preserving Clustering

Technique Using Hybrid Data Transformation Method.” in In Proc. IEEE International Conference, 2009, PP. 1502 - 1506.

[11] Mohammad Ali Kadampur, D.V.L.N Somayajulu, S.S. Shivaji Dhiraj, and Shailesh G.P. Satyam, “Privacy preserving clustering by cluster bulging for information sustenance.” In Proc. 4th International Conference on Information and Automation for Sustainability (ICIAfS), Colombo, Sri Lanka, 2008, pp. 158-164.

[12] Jie Liu, Yifeng XU, Harbin, “privacy preserving clustering by Random Response Method of Geometric Transformation.” In Proc. Fourth international conference on internet computing for science and engineering, 2009, pp. 181-188.

[13] Keke Chen, Ling Liu, “Geometric data perturbation for privacy preserving outsourced data mining.” Knowledge and Information Systems journal, Volume 29, Issue 3, pp 657-695, December 2011.

[14] Khaled Alotaibi, V. J. Rayward-Smith, Wenjia Wang, and Beatriz de la Iglesia, “Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification.” in Proc. ASE/IEEE International Conference on Social Computing, 2012 and ASE/IEEE International Conference on Privacy, Security, Risk and Trust, 2012, pp. 694 - 701.

[15] Ms Shalini Lamba, Dr S. Qamar Abbas, “A model for preserving privacy of sensitive data.” International Journal of Technical Research and Applications Vol. 1, No. 3, PP. 07-11, 2013.

[16] MohammadReza Keyvanpour, Somayyeh Seifi Moradi, “Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework.” International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 2, Feb 2011.

[17] Keerti Dixit, Bhupendra Pandya, “An overview of Multiplicative data perturbation for privacy preserving Data mining.” International Journal for research in applied science and engineering technology (I JRAS ET), Vol. 2, Issue VII, pp 90-96, July 2014.

[18] CHEN, K., and LIU, L. “A random rotation perturbation approach to privacy preserving data classification.” In Proc. Intl. Conf. on Data Mining (ICDM) 2005.

[19] http://www.cs.waikato.ac.nz/ml/weka/

Leila Jafar Tafreshi is a MSc graduate of Artificial Intelligence at Semnan University in 2015. She received her BSc degree from Kharazmi University in 2012. Her research interests include Privacy and Data mining.

Farzin Yaghmaee received his PhD in 2010 and MSc in 2002 both in Artificial Intelligence from Sharif University of Technology, Iran and received BSc from AmirKabir University of Technology. He is now a faculty member of Electrical and Computer Engineering Department of Semnan University. His research interests are: image and video processing, text mining and Persian language processing tools.

Promote Mobile Banking Services by using National Smart Card Capabilities and NFC Technology

Reza Vahedi*
Department of IT Management, Electronic Branch, Islamic Azad University, Tehran, Iran
it.vahedi@yahoo.com
Farhad Hosseinzadeh Lotfi
Department of Mathematics,Science and Research Branch, Islamic Azad University, Tehran, Iran
farhad@hosseinzadeh.ir
Seyed Esmaeial Najafi
Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
najafiisis@yahoo.com

Received: 06/Jun/2016 Revised: 03/Aug/2016 Accepted: 13/Aug/2016

Abstract

By the mobile banking system and install an application on the mobile phone can be done without visiting the bank and at any hour of the day, get some banking operations such as account balance, transfer funds and pay bills did limited. The second password bank account card, the only security facility predicted for use mobile banking systems and financial transactions. That this alone cannot create reasonable security and the reason for greater protection and prevent the theft and misuse of citizens' bank accounts is provide banking services by the service limits. That by using NFC (Near Field Communication) technology can identity and biometric information and Key pair stored on the smart card chip be exchanged with mobile phone and mobile banking system. And possibility of identification and authentication and also a digital signature created documents. And thus to enhance the security and promote mobile banking services. This research, the application and tool library studies and the opinion of seminary experts of information technology and electronic banking and analysis method Dematel is examined. And aim to investigate possibility Promote mobile banking services by using national smart card capabilities and NFC technology to overcome obstacles and risks that are mentioned above. Obtained Results, confirmed the hypothesis of the research and show that by implementing the so-called solutions in the banking system of Iran.

Keywords: NFC Technology; National Smart Card; Mobile Banking; Identity; Security.

1. Introduction

Today, by the mobile banking system and install an application on the mobile phone can be done without visiting the bank and at any hour of the day, get some banking operations such as account balance, transfer funds and pay bills did limited [1]. The second password bank account card, the only security facility predicted for use mobile banking systems and financial transactions. That this alone cannot create reasonable security and the reason for greater protection and prevent the theft and misuse of citizens' bank accounts is provide banking services by the service limits.

With the expanding use of smart phones and add NFC technology to this type of phones Applications and new capabilities in the tech world has been created That Comfort and increase the speed and security of different activities, such as Share files, used in opening and closing the doors locked, read information NFC tags installed on the books in the library, etc. [2].

The purpose of this research, exploring the possibility of promoting mobile banking services by using national smart card capabilities and NFC that is a new technology [3,4]. That this goal is by inserting the national smart card alongside mobile and Creation wireless communicate between them by the NFC technology for exchange information stored in the national smart card chip (Identity information, biometrics and digital signing keys) with the mobile banking system, it is possible [5]. And increase the level of security and thus enabling the development and promoting mobile services offered by banks.

1.1 History and Research Literature

In this section an overview of the history and technology concepts NFC, national smart card, mobile banking as well as the dematel method will have.

1.1.1 National Smart Card

National smart identity card (e_ID Card): The New Generation card having an electronic chip that can be programmed like a computer and have the memory, processor, operating system and application installation are applets. Issued by the organization for Civil registration and containing identity information and biometric (fingerprints and facial image) with the possibility of issuing identity certificates, digital signatures and authentication biometric (MOC) to identify and authenticate citizens [6,7].

* Corresponding Author

Smart cards can be either contact or contactless smart card. Smart cards may provide strong security authentication for single sign-on (SSO) within large organizations.

Smart card types

- Contact Memory Card
- Contact CPU Card
- Contact-less Memory Card
- Dual Interface CPU Card

In national smart ID card project uses of type dual Interface CPU card. In the type of cards implement contactless and contact interfaces on a single card with some shared storage and processing.

Design

A smart card may have the following generic characteristics:

- Similar dimensions to those of a credit card. ID-1 of the ISO/IEC 7810 standard defines cards as nominally 85.60 by 53.98 millimeters (3.370 in × 2.125 in). Another popular size is ID-000 which is nominally 25 by 15 millimeters (0.984 in × 0.591 in) (commonly used in SIM cards). Both are 0.76 millimeters (0.030 in) thick.
- Contains a tamper-resistant security system (for example a secure cryptoprocessor and a secure file system) and provides security services (e.g., protects in-memory information).
- Managed by an administration system which securely interchanges information and configuration settings with the card, controlling card blacklisting and application-data updates.

Development of contactless systems

Contactless smart cards do not require physical contact between a card and reader. They are becoming more popular for payment and ticketing. Typical uses include mass transit and motorway tolls. Visa and MasterCard implemented a version deployed in 2004–2006 in the U.S. Most contactless fare collection systems are incompatible, though the MIFARE Standard card from NXP Semiconductors has a considerable market share in the US and Europe.

Smart cards are also being introduced for identification and entitlement by regional, national, and international organizations. These uses include citizen cards, drivers' licenses, and patient cards.

Smart card applications in finance:

Smart cards serve as credit or ATM cards, fuel cards, mobile phone SIMs, authorization cards for pay television, household utility pre-payment cards, high-security identification and access-control cards, and public transport and public phone payment cards.

Smart cards may also be used as electronic wallets. The smart card chip can be "loaded" with funds to pay parking meters, vending machines or merchants. Cryptographic protocols protect the exchange of money between the smart card and the machine. No connection to a bank is needed. The holder of the card may use it even if not the owner. Examples are Proton, Geldkarte, Chipknip and Moneo. The German Geldkarte is also used to validate customer age at vending machines for cigarettes.

Smart card applications in Identification

Smart-cards can authenticate identity. Sometimes they employ a public key infrastructure (PKI). The card stores an encrypted digital certificate issued from the PKI provider along with other relevant information. Examples include the U.S. Department of Defense (DoD) Common Access Card (CAC), and other cards used by other governments for their citizens. If they include biometric identification data, cards can provide superior two- or three-factor authentication.

Smart cards are not always privacy-enhancing, because the subject may carry incriminating information on the card. Contactless smart cards that can be read from within a wallet or even a garment simplify authentication; however, criminals may access data from these cards.

Cryptographic smart cards

Cryptographic smart cards are often used for single sign-on. Most advanced smart cards include specialized cryptographic hardware that uses algorithms such as RSA and Digital Signature Algorithm (DSA). Today's cryptographic smart cards generate key pairs on board, to avoid the risk from having more than one copy of the key (since by design there usually isn't a way to extract private keys from a smart card). Such smart cards are mainly used for digital signatures and secure identification.

The most common way to access cryptographic smart card functions on a computer is to use a vendor-provided PKCS#11 library. [citation needed] On Microsoft Windows the Cryptographic Service Provider (CSP) API is also supported.

The most widely used cryptographic algorithms in smart cards (excluding the GSM so-called "crypto algorithm") are Triple DES and RSA. The key set is usually loaded (DES) or generated (RSA) on the card at the personalization stage.

Some of these smart cards are also made to support the National Institute of Standards and Technology (NIST) standard for Personal Identity Verification, FIPS 201.

Advantages

The first main advantage of smart cards is their flexibility. Smart cards have multiple functions which simultaneously can be an ID, a credit card, a stored-value cash card, and a repository of personal information such as telephone numbers or medical history. The card can be easily replaced if lost, and, the requirement for a PIN (or other form of security) provides additional security from unauthorised access to information by others. At the first attempt to use it illegally, the card would be deactivated by the card reader itself.

The second main advantage is security. Smart cards can be electronic key rings, giving the bearer ability to access information and physical places without need for online connections. They are encryption devices, so that the user can encrypt and decrypt information without relying on unknown, and therefore potentially untrustworthy, appliances such as ATMs. Smart cards are very flexible in providing authentication at different level of the bearer and the counterpart. Finally, with the information about the user that smart cards can provide to the other parties, they are useful devices for customizing products and services.

Other general benefits of smart cards are:

- Portability
- Increasing data storage capacity
- Reliability that is virtually unaffected by electrical and magnetic fields.

Smart cards and electronic commerce

Smart cards can be used in electronic commerce, over the Internet, though the business model used in current electronic commerce applications still cannot use the full potential of the electronic medium. An advantage of smart cards for electronic commerce is their use customize services. For example, in order for the service supplier to deliver the customized service, the user may need to provide each supplier with their profile, a boring and time-consuming activity. A smart card can contain a non-encrypted profile of the bearer, so that the user can get customized services even without previous contacts with the supplier [4,6,7].

1.1.2 Mobile Banking:

Mobile banking is a service provided by a bank or other financial institutionthat allows its customers to conduct a range of financial transactionsremotely using a mobile device such as a mobile phone or tablet, and using software, usually called an app, provided by the financial institution for the purpose. Mobile banking is usually available on a 24-hour basis. Some financial institutions have restrictions on which accounts may be accessed through mobile banking, as well as a limit on the amount that can be transacted.

The types of financial transactions which a customer may transact through mobile banking include obtaining account balances and list of latest transactions, electronic bill payments, and funds transfers between a customer's or another's accounts. Some also enable copies of statements to be downloaded and sometimes printed at the customer's premises; and some banks charge a fee for mailing hardcopies of bank statements.

From the bank's point of view, mobile banking reduces the cost of handling transactions by reducing the need for customers to visit a bank branch for non-cash withdrawal and deposit transactions. Transactions involving cash or documents (such as cheques) are not able to be handled using mobile banking, and a customer needs to visit an ATM or bank branch for cash withdrawals and cash or cheque deposits.

Mobile banking differs from mobile payments, which involves the use of a mobile device to pay for goods or services either at the point of sale or remotely, analogously to the use of a debit or credit card to effect an EFTPOS payment.

History

The earliest mobile banking services used SMS, a service known as SMS banking. With the introduction of smart phoneswith WAP support enabling the use of the mobile web in 1999, the first European banks started to offer mobile banking on this platform to their customers. Mobile banking has until recently (2010) most often been performed via SMS or the mobile web. Apple's initial success with iPhone and the rapid growth of phones based on Google's Android (operating system) have led to increasing use of special client programs, called apps, downloaded to the mobile device. With that said, advancements in web technologies such as HTML5, CSS3 and JavaScript have seen more banks launching mobile web based services to complement native applications. A recent study (May 2012) by Mapa Research suggests that over a third of banks have mobile device detection upon visiting the banks' main website. A number of things can happen on mobile detection such as redirecting to an app store, redirection to a mobile banking specific website or providing a menu of mobile banking options for the user to choose from.

A mobile banking conceptual

In one academic model, mobile banking is defined as:

Mobile Banking refers to provision and availment of banking- and financial services with the help of mobile telecommunication devices. The scope of offered services may include facilities to conduct bank and stock market transactions, to administer accounts and to access customised information.

According to this model mobile banking can be said to consist of three inter-related concepts:

- Mobile accounting
- Mobile brokerage
- Mobile financial information services

Most services in the categories designated accounting and brokerage are transaction-based. The non-transaction-based services of an informational nature are however essential for conducting transactions - for instance, balance inquiries might be needed before committing a money remittance. The accounting and brokerage services are therefore offered invariably in combination with information services. Information services, on the other hand, may be offered as an independent module. Mobile banking may also be used to help in business situations as well as financial.

Mobile banking services

Typical mobile banking services may include:

- Account - information
- Transaction
- Investments
- Support
- Content services

A report by the US Federal Reserve found that 21 percent of mobile phone owners had used mobile banking in the past 12 months.[5] Based on a survey conducted by Forrester, mobile banking will be attractive mainly to the younger, more "tech-savvy" customer segment. A third of mobile phone users say that they may consider performing some kind of financial transaction through their mobile phone. But most of the users are interested in performing basic transactions such as querying for account balance and making bill payment.

Security

As with most internet- connected devices, as well as mobile-telephony devices, cybercrime rates are escalating year-on-year. The types of cybercrimes which may affect mobile-banking might range from unauthorized use while the owner is using the toilet, to remote-hacking, or even jamming or interference via the internet or telephone network data streams. In the banking world, currency rates may change by the millisecond.

Security of financial transactions, being executed from some remote location and transmission of financial information over the air, are the most complicated challenges that need to be addressed jointly by mobile application developers, wireless network service providers and the banks' IT departments.

The following aspects need to be addressed to offer a secure infrastructure for financial transaction over wireless network:

1. Physical part of the hand-held device. If the bank is offering smart-card based security, the physical security of the device is more important.
2. Security of any thick-client application running on the device. In case the device is stolen, the hacker should require at least an ID/Password to access the application.
3. Authentication of the device with service provider before initiating a transaction. This would ensure that unauthorized devices are not connected to perform financial transactions.
4. User ID / Password authentication of bank’s customer.
5. Encryption of the data being transmitted over the air.
6. Encryption of the data that will be stored in device for later / off-line analysis by the customer.

One-time password (OTPs) are the latest tool used by financial and banking service providers in the fight against cyber fraud. [8] Instead of relying on traditional memorized passwords, OTPs are requested by consumers each time they want to perform transactions using the online or mobile banking interface. When the request is received the password is sent to the consumer’s phone via SMS. The password is expired once it has been used or once its scheduled life-cycle has expired.

Because of the concerns made explicit above, it is extremely important that SMS gateway providers can provide a decent quality of service for banks and financial institutions in regards to SMS services. Therefore, the provision of service level agreements (SLAs) is a requirement for this industry; it is necessary to give the bank customer delivery guarantees of all messages, as well as measurements on the speed of delivery, throughput, etc. SLAs give the service parameters in which a messaging solution is guaranteed to perform.

Mobile banking in the world

Mobile banking is used in many parts of the world with little or no infrastructure, especially remote and rural areas. This aspect of **mobile commerce** is also popular in countries where most of their population is **unbanked**. In most of these places, banks can only be found in big cities, and customers have to travel hundreds of miles to the nearest bank.

In Iran, banks such as Parsian, Tejarat, Pasargad Bank, Mellat, Saderat, Sepah, Edbi, and Bank melli offer the service [1].

1.1.3 NFC Technology

Near Field Communication or NFC technology is a short-range wireless encrypted communication at a distance of 4 cm or less that in the frequency band of 13.56 MHz ability to exchange data with speed 424 KB / s (on average). NFC can be with contactless smart cards ISO / IEC 1443 available as well as other devices equipped with the technology to easily communicate and exchange information with them. NFC specifically designed to work on mobile devices and has three general features’ that make it transparent development process. In the first feature, this technology has the potential to be used instead of the existing contactless cards so much so that you can use exactly like cards available for micropayments. In the second feature, you can use this technology as a reader and RFID passive tags and use it in promotional interactivity. The third feature of this technology also this capability gives you that as a reader and sender used this feature and in a state person-to-person exchange of information between two powered device NFC to take advantage of it [2].

Applications

NFC allows one- and two-way communication between endpoints, suitable for many applications.

Commerce

NFC devices can be used in contactless payment systems, similar to those used in credit cards and electronic ticket smartcards and allow mobile payment to replace/supplement these systems.

In Android 4.4, Google introduced platform support for secure NFC-based transactions through Host Card Emulation (HCE), for payments, loyalty programs, card access, transit passes and other custom services. HCE allows any Android 4.4 app to emulate an NFC smart card, letting users initiate transactions with their device. Apps can use a new Reader Mode to act as readers for HCE cards and other NFC-based transactions.

On September 9, 2014, Apple announced support for NFC-powered transactions as part of Apple Pay. Apple stated that their approach to NFC payment is more secure because Apple Pay tokenizes its data to encrypt and protect it from unauthorized use.

Bootstrapping other connections

NFC offers a low-speed connection with simple setup that can be used to bootstrap more capable wireless connections. For example, Android Beam software uses NFC to enable pairing and establish a Bluetooth connection when doing a file transfer and then disabling Bluetooth on both devices upon completion. Nokia, Samsung, BlackBerry and Sony have used NFC technology to pair Bluetooth headsets, media players and speakers with one tap. [citation needed] The same principle can be applied to the configuration of Wi-Fi networks. Samsung Galaxy devices have a feature named S-Beam—an extension of Android Beam that uses NFC (to share MAC Address and IP addresses) and then uses Wi-Fi Direct to share files and documents. The advantage of using Wi-Fi Direct over Bluetooth is that it permits much faster data transfers, running up to 300Mbit/s.

Social networking

NFC can be used for social networking, for sharing contacts, photos, videos or files and entering multiplayer mobile games.

Identity and access tokens

NFC-enabled devices can act as electronic identity documents and keycards.[53] NFC's short range and encryption support make it more suitable than less private RFID systems.

Smartphone automation and NFC tags

NFC-equipped smartphones can be paired with NFC Tags or stickers that can be programmed by NFC apps. These programs can allow a change of phone settings, texting, app launching, or command execution.

Such apps do not rely on a company or manufacturer, but can be utilized immediately with an NFC-equipped smartphone and an NFC tag.

The NFC Forum published the Signature Record Type Definition (RTD) 2.0 in 2015 to add integrity and authenticity for NFC Tags. This specification allows an NFC device to verify tag data and identify the tag author.

1.1.4 Dematel Technique

Dematel technique commonly used to investigate the very complex issues and apply for structuring a sequence of Supposed information. So that the intensity of relationship to be examined scoring, coupled with the importance of their Feedback to make search and accepts non-transferable relationships [8].

Dematel technique has two major functions:

1. Considering the mutual communication; advantage of this method compared to network analysis technique, clarity it to reflect mutual communication is among a wide range of components. So that specialists are able with greater mastery to express their opinions about the effects, the direction and severity of affective between the factors them.

2. Structuring the complex factors in groups of cause and effect. This case is one of the most important functions and one of the most important reasons for its frequent application in problem-solving processes. So that the classification of a wide range of complex factors in the form of cause-effect, the decision maker in better condition to understand the relationships. This results in recognizing the crucial status and role in the process of mutual effecting.

1.1.5 Research History

To some of the research on mobile banking, NFC and their applications and also the use of tools dematel for analyzing data in the studies, will be mentioned in table1.

Table 1. Research history

Year	Researcher	Work done	Ref. No
2015	Mohammadi S, Barkhordari Firouzabadi M	They offer a model for authentication and admission patient & payment of fees by health smart card, NFC and mobile payments	[9]
2013	Hosseinpour Soleymani A, Yousefi M	They assessed application and risk, mobile payment transactions by NFC technology	[10]
2014	Mohammadzadeh A, Ataei M, Salimi H	Identify and prioritize barriers the collected retarded banking debts using a combination model Dematel network and vikor	[11]
2013	Tabatabaei M, Hosseini S, Noori A	Identify and prioritize the criteria of quality services MCDM approach in the banking industry (by dematel technique)	[12]

2. Doing this Study Method and Tools for Data Collection and Analysis:

This research, using library studies and specialists, IT & electronic banking and interviews with them, possibility promoting mobile banking services using the capabilities of the national smart card and NFC technology review and after gathering the required information the mentioned method and criteria for specify and after preparing the appropriate paired comparisons questionnaire and complete it by the experts, analyzed the results using Dematel process.

3. Research Methodology

In figure 1 steps of this research are shown separately:

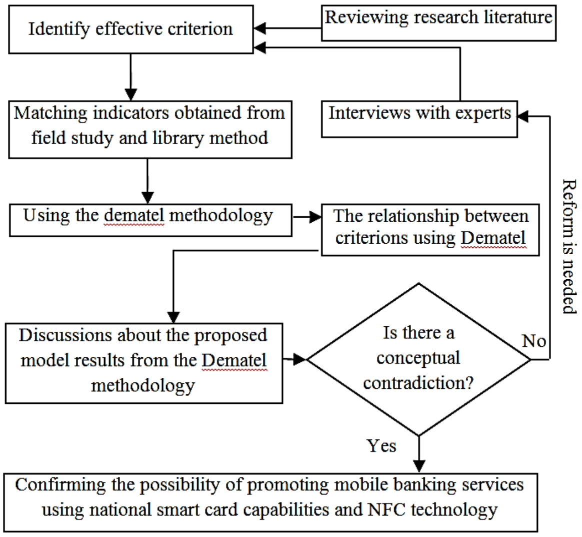

Identify effective criterion
Reviewing research literature
Matching indicators obtained from field study and library method
Interviews with experts
Using the dematel methodology
The relationship between criterions using Dematel
Reform is needed
Discussions about the proposed model results from the Dematel methodology
Is there a conceptual contradiction?
No
Yes
Confirming the possibility of promoting mobile banking services using national smart card capabilities and NFC technology

Fig 1. Research steps

3.1 Determine the effective criteria

According to the research literature and expert opinion, an effective criterion to achieve the research objectives within the framework of six criteria in table 2 was determined.

Table 2. Effective criteria

Proprietary word	Criteria	Definition
C_1	Identification	Identify the people, by self-expression profile and provide identification documents
C_2	Authentication	Proof of correctness, identity of people information that was given in Identification stage
C_3	Undeniable signature	Acceptance and approval documents, commitments and requests in electronic transactions so that there is no possibility of denial.
C_4	Security requirements	Protocols and security guidelines
C_5	Technical Requirements	Software and equipment, hardware and communication Infrastructure
C_6	Macro banking services	Main banking services such as account opening, apply for a loan, transfer too much money and...

Applicability and purpose of this study was determined as follows:

A (application): use of national smart card capabilities in mobile banking system by NFC technology

P (main purpose): promoting security and development services provided by mobile banking.

3.2 Analysis of relations among of criteria and create a total relation matrix using Dematel

The structure decision to confirm possibility promote mobile banking services by using national smart card capabilities and NFC technology were discussed and reviewed conforms to the figure 2, According to which and based on dematel methodology, domestic relationships are managed.

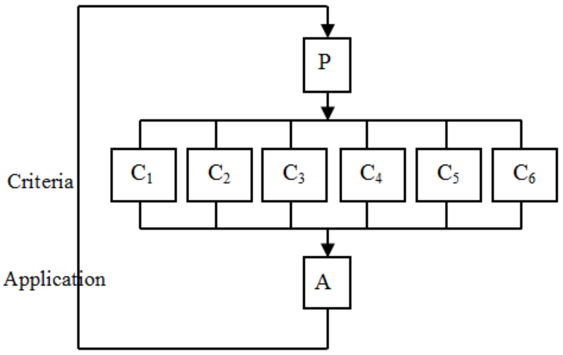


Fig. 2. Structure Design

To determine the relationship between effective criteria, dematel technique is used. This technique involves four main steps. In the first stage should be paired comparison matrix of criteria that were scored by experts and its numbers are ranging from 0 to 4, to be created. The number zero is showing that they are on each other ineffective and number 4 is representing the greatest impact. To review criteria, from the standpoint of five experts has been used that for the consideration of the opinion of all experts, according to formula 1 of them we the arithmetic mean.

$$Z = \frac{x^1+x^2+x^3+\cdots+x^p}{p} \qquad (1)$$

P In this formula, the number of experts and x^1, x^2... x^p, respectively, are paired comparison matrix expert 1, expert 2 and expert p and for normalizing the matrix obtained from the formulas 2 and 3 is used.

$$H_{ij} = \frac{z_{ij}}{r} \qquad (2)$$

That r is obtained from the following formula:

$$r = \max_{1\le i\le n}\left(\sum_{j=1}^{n} z_{ij}\right) \qquad (3)$$

Table 3 shows the normalized matrix.

Table 3. Normalized matrix

	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0.00	0.19	0.21	0.19	0.13	0.21
C_2	0.00	0.00	0.24	0.26	0.21	0.29
C_3	0.00	0.00	0.00	0.28	0.26	0.29
C_4	0.00	0.00	0.00	0.00	0.21	0.26
C_5	0.00	0.00	0.00	0.00	0.00	0.18
C_6	0.00	0.00	0.00	0.00	0.00	0.00

After calculating the above matrix, the total relation matrix is obtained according to the formula 4.

$$T = \lim_{k\to+\infty}\left(H^1 + H^2 + \cdots + H^k\right) = H \times (I - H)^{-1} \qquad (4)$$

In this formula, *I* is the unit matrix.

Table 4 shows the T matrix.

Table 4. Total relation matrix

	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0.00	0.19	0.25	0.31	0.30	0.47
C_2	0.00	0.00	0.24	0.33	0.34	0.51
C_3	0.00	0.00	0.00	0.28	0.32	0.42
C_4	0.00	0.00	0.00	0.00	0.21	0.30
C_5	0.00	0.00	0.00	0.00	0.00	0.18
C_6	0.00	0.00	0.00	0.00	0.00	0.00

The next step is to obtain the sum of rows and columns of the matrix T. The sum of rows and columns according to formulas 5 and 6 obtained.

$$(D)_{n\times 1} = \left[\sum_{j=1}^{n} T_{ij}\right]_{n\times 1} \qquad (5)$$

$$(R)_{1\times n} = \left[\sum_{j=1}^{n} T_{ij}\right]_{1\times n} \qquad (6)$$

That D and R are respectively matrix n × 1 and 1 × n.

Next, the importance of (D_i + R_i) and the relationship between the criteria (D_i-R_i) is determined. If D_i-R_i>0 is the relevant criteria is effective and if D_i-R_i<0 is the relevant criteria is receptive effect. Table 5 D_i+R_i and D_i-R_i shows.

Table 5. The importance and effectiveness criteria

Criteria	$D_i + R_i$	$D_i - R_i$
Criterion 1	1.53	1.53
Criterion 2	1.60	1.22
Criterion 3	1.51	0.54
Criterion 4	1.43	-0.41
Criterion 5	1.34	-0.99
Criterion 6	1.88	-1.88

Chart 1 shows the importance and effectiveness between the criteria. The horizontal axis shows the importance of the criteria and the vertical axis shows the impact or receptive effect the criteria.

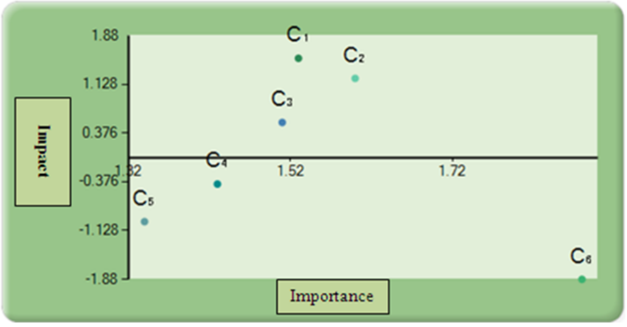


Chart 1. The relationships and the importance of criteria

According to the Chart 1, the criteria C_1 (identification), C_2 (authentication), C_3 (signature undeniable), effective criteria (cause) and the criteria C_4 (security requirements), C_5 (technical requirements), C_6 (macro banking services) as criteria affected (disabled), are introduced.

In the final step, according to negotiate with experts, threshold value in this study, average total numbers obtained from the matrix table of the total relationships (direct and indirect relationships) were considered. Therefore, this study is a threshold value equal 0.14.

On this basis and according to the results of total relation matrix, As shown in Figure 3, criterion C_1 to C_2, C_3, C_4, C_5, C_6 and criterion C_2 to C_3, C_4, C_5, C_6 and criterion C_3 to C_2, C_4, C_5, C_6 and criterion C_4 to C_5, C_6 and criterion C_5 to C_6 is affected. And to criterion C_6 (macro banking services) affect all other criteria. In other words criteria C_1 up to C_5 affecting the criterion C_6 (Macro Banking Services) is.

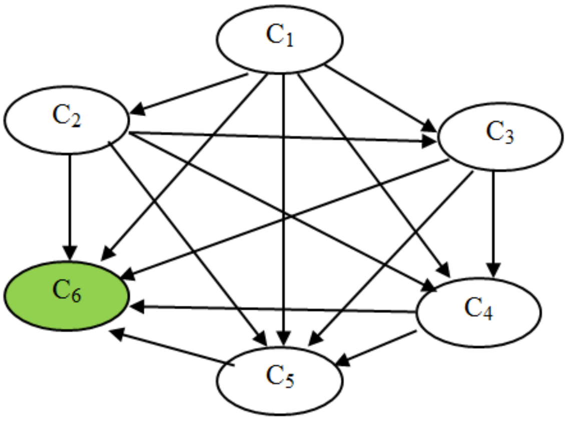


Fig. 3. Relationship between criteria

4. Conclusions

The aim of this study was to promote mobile banking services by using the capabilities of the National Smart Card and NFC technology. That's according to the research literature and expert opinion and the results of data analysis research by providing conditions (criteria C_1 to C_6) by inserting the national smart card alongside mobile and Creation wireless communicate between them by the NFC technology for exchange information stored in the national smart card chip (Identity information, biometrics and digital signing keys) with the mobile banking system, it is possible. And increase the level of security and thus enabling the development and promoting mobile services offered by banks.

4.1 Future research

It is suggested to the researchers that the use of the national smart ID card and NFC technology in the field of micro-payments, investigate.

References

[1] https://en.wikipedia.org/wiki/Mobile_banking

[2] https://en.wikipedia.org/wiki/Near_field_communication

[3] Project Management Office, National smart card applications, Tehran: NOCR, 2010, pp.5-45.

[4] Project Management Office, Some of the dimensions of a national smart card application, Tehran: NOCR, 2010, pp.12-67.

[5] A. Riyazi, National Smart Card (Secure, standards, biometrics), Tehran: Nass, 2009, pp.63-101.

[6] NOCR Education Center, The National Smart Card-structure and its role in e-government, Tehran: Institute of Danesh Parsiyan, 2011, pp.51-64.

[7] Project Management Office, National Smart Card Project for Metropolitan Architecture, Tehran: NOCR, 2010, pp.167-198.

[8] S. M. Arabi, and N. Azad, "The effects of the implementation of the ASYCUDA system in the country's business sector", Iranian journal of Trade Studies, Vol. 29, pp.137-165, 2003.

[9] M. Barkhordari Firouzabadi, and S. Mohammadi, "A Review and Proposal of a Model for Patient Authentication and NFC Mobile Payment with Smart Health Card", in 1st National E-Conference of Technology Developments on Electrical, Electoroncs and Computer Engineering, Iran, 2015.
[10] A. Hosseinpour Soleymani, and M. Yousefi," Introduction to NFC technology and its application in mobile payments", in the 1st Regional Conference on Information Technology, Iran, 2013.
[11] A. Mohammadzadeh, and M. Ataei, and H. Salimi, "Identifying and Prioritizing the Obstacles Leading to Bank Overdue, Using DEMATEL and VIKOR", Journal of Development Evolution Management, Vol.16, pp.15-26, 2014.
[12] M. Tabatabaei, and S, Hosseini, and A. Noori, " Identify and prioritize the criteria of quality services MCDM approach in the banking industry", in 9th International Conference on Industrial Engineering, Iran, 2013.

Reza Vahedi is a M.A graduate of Information Technology Management at Electronic Branch, Islamic Azad University. He got his B.Sc in Computer Engineering. His research interests include Business intelligence (BI), decision support system (DSS), Management Information System (MIS), software Engineering, Expert systems (ES), Data modeling And e-banking.

Farhad Hosseinzadeh Lotfi is a Ph.D graduate of Applied Mathematics (O. R.) at Science & Research Branch, Islamic Azad University, Tehran in 2000. He currently is a professor of Operation Research. His main research interests are in the fields of Operation Research, Data Envelopment Analysis (DEA) and Application of DEA in Banking. He has published more than 340 papers in various journals and conferences as well as 17 books.

Sayed Esmaeail Najafi is a Ph.D graduate of Industrial Engineering(System Management and efficiency) at Science and Research Branch, Islamic Azad University, Tehran in 2009. His research in the fields of Data Envelopment Analysis (DEA), organization Architecture and Decision Making Methods. He has published more than 31 papers in various journals and conferences as well as 5 books.

Data Aggregation Tree Structure in Wireless Sensor Networks Using Cuckoo Optimization Algorithm

Elham Mohsenifard
Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran
mohsenifards@gmail.com
Ali Ghaffari*
Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran
a.ghaffari@iaut.ac.ir

Received: 28/May/2016 Revised: 25/Jul/2016 Accepted: 02/Aug/2016

Abstract

Wireless sensor networks (WSNs) consist of numerous tiny sensors which can be regarded as a robust tool for collecting and aggregating data in different data environments. The energy of these small sensors is supplied by a battery with limited power which cannot be recharged. Certain approaches are needed so that the power of the sensors can be efficiently and optimally utilized. One of the notable approaches for reducing energy consumption in WSNs is to decrease the number of packets to be transmitted in the network. Using data aggregation method, the mass of data which should be transmitted can be remarkably reduced. One of the related methods in this approach is the data aggregation tree. However, it should be noted that finding the optimization tree for data aggregation in networks with one working-station is an NP-Hard problem. In this paper, using cuckoo optimization algorithm (COA), a data aggregation tree was proposed which can optimize energy consumption in the network. The proposed method in this study was compared with genetic algorithm (GA), Power Efficient Data gathering and Aggregation Protocol- Power Aware (PEDAPPA) and energy efficient spanning tree (EESR). The results of simulations which were conducted in matlab indicated that the proposed method had better performance than GA, PEDAPPA and EESR algorithm in terms of energy consumption. Consequently, the proposed method was able to enhance network lifetime.

Keywords: Wireless Sensor Networks (WSNs); Data Aggregation Technique; Data Aggregation Tree; Cuckoo Optimization Algorithm (COA); Network Lifetime Enhancement.

1. Introduction

WSNs are resource constrained networks and include many limited-energy sensor nodes. In WSN, each sensor can sense specific data and transmit it to its neighbors [1]. In general, a central node, namely sink is the destination of all data packets. For transmitting data to long distances, a lot of energy should be consumed. Hence, in many cases, nodes communicate with the sink node through their neighbors. In this case, each node should know which neighbor is more appropriate for packet transmission. In WSNs, congestion control mechanisms are important techniques for decreasing timeliness and increasing packet delivery rate [2].

Communication protocols play a significant role in enhancing the efficiency and lifetime of WSNs [3,4]. On the other hand, energy is one of the most important parameter and the key objectives in data gathering in WSNs. Hence, designing efficient protocols with regard to energy consumption for WSNs is essential since it can not only reduce the total energy consumption in the network but also distribute energy steadily and uniformly to the network nodes. Recently, many algorithms have been proposed for data aggregation in WSNs which try to find routes towards the sink through which data can be aggregated.

Given a data aggregation tree, sensors receive messages from children periodically, merge them with its own packet, and send the new packet to its parent. The problem of finding an aggregation tree with the maximum lifetime has been proved to be NP-hard and can be generalized to finding a spanning tree with the minimum maximum vertex load, where the load of a vertex is a nondecreasing function of its degree in the tree [5].

Among the proposed protocols, data aggregation technique [6], [7] is an energy conservation scheme which tries to decrese the volume of data communicated by collecting local data at intermediate nodes and forwarding only the result of an aggregation operation, such as min and max, towards the sink node [8], [9], [10]. In data aggregation technique [11], [12], [13], [14], [15] relevant data packets were combined with one another in intermediate nodes and form a packet. Hence, the number of packets to be transmitted in the network decreases [16], [17]. Consequently, less energy will be consumed [16].

In this paper, an efficient and low-energy data aggregation method based on COA [18] is introduced which can reduce the total communication energy through creating a data aggregation tree. The problem addressed in this paper is an important special case of the general data aggregation problem in which all the sensor nodes in the network are source nodes. Hence, it can efficiently

* Corresponding Author

reduce energy consumption and enhance network lifetime. The objective in this paper is to simultaneously minimize total energy consumption during the entire tree structure.

The rest of the paper is organized as follows: in section 2, a summary of the related studies and works is briefly overviewed. Then, the proposed algorithm is described in section 3. Then, simulation results are reported in section 4. Finally, the conclusion of the study is given in section 5.

2. Related Works

Regarding related studies on data aggregation, In [19], PEDAP (Power Efficient Data Gathering and Aggregation Protocol) was proposed. PEDAP computes a routing tree with MST (Minimum Spanning Tree) algorithm by setting the energy cost delivering a data packet between two sensors as the weight of the link between them. However, routing trees constructed with MST are not optimal energy efficient routing trees and cannot obtain long network lifetime. Based on PEDAP, another algorithm, PEDAP-PA (Power Efficient Data Gathering and Aggregation Protocol-Power Aware) [19], was proposed two minimum spanning tree based data gathering and aggregation schemes to maximize the lifetime of the network, where one is the power aware version of the other. The non-power aware version (PEDAP) extends the lifetime of the last node by minimizing the total energy consumed from the system in each data gathering round, while the power aware version (PEDAPPA) balances the energy consumption among nodes. In PEDAP, edge cost is computed as the sum of transmission and receiving energy. In PEDAPPA, however, an asymmetric communication cost is considered by dividing PEDAP edge cost with transmitter residual energy. A nodewith higher edge cost is included later in the tree which results few incoming messages. Once the edge cost is established, routing information is computed using Prim’s minimum spanning tree rooted at base station. The routing information is computed periodically after a fixed number of rounds. These algorithms assume all nodes perform in-network data aggregation and base station is aware of the location of the nodes.

In [20], EESR is proposed that uses energy efficient spanning tree based multi-hop to extend network lifetime. EESR generates a transmission schedule that contains a collection of routing trees. In EESR, the lowest energy node is selected to calculate its edge cost. A node calculates its outgoing edge cost by taking the lower energy level between sender and receiver. Next, the highest edge cost link is chosen to forward data towards base station. If the selected link creates a cycle, then the next highest edge cost link is chosen. This technique avoids a node to become overloaded with too many incoming messages. The total procedure is repeated for the next lowest energy node. As a result, higher energy nodes calculate their edge costs later and receive more incoming messages, which balances the overall load among nodes. The protocol also generates a smaller transmission schedule, which reduces receiving energy. The base station may broadcast the entire schedule or an individual tree to the network.

Wang et al [21] devised a data aggregation tree based on first-fit algorithm for reducing data transmission delay. Chaon et al [22] proposed SFEB (structure-free and energy-balanced) protocol which is an unstructured data aggregation protocol; it results in a balanced energy consumption among the nodes. Consequently, network lifetime is enhanced. Liu et al [23] proposed an approximation algorithm using directional antennas for aggregating data which led to the reduction of interface in transmitting and receiving data in network. Using ant colony algorithm, Liao et al [24] proposed a routing protocol which increased network lifetime. Islam et al., [25] produced a balanced data aggregation tree based on genetic algorithm which was energy-efficient. The authors used GA to create multi-hop spanning trees and it was not based on hierarchical clusters. The data aggregation trees created by the proposed GA technique were more energy efficient than some of the current data aggregation tree-based techniques.

Using lexicographic method, Lai and Ravindran [26] investigated accessing maximum network lifetime and fairness regarding simultaneous resource allocation in data aggregation tree. Kwond et al., [27] allocated a dynamic time for data aggregation in a node which was characterized with high energy efficiency and little delay overhead resistance. Al-Karaki et al [28] proposed a design which used optimization and heuristic algorithms simultaneously for locating minimum data aggregation points and routing data to base station so that network lifetime is maximized.

In [29], Firstly, the authors proposed an adaptive spanning tree algorithm (AST), which can adaptively build and adjust an aggregation spanning tree. Owing to the strategies of random waiting and alternative father nodes, AST can achieve a relatively balanced spanning tree and flexible tree adjustment. Then, the authors presented a redundant aggregation scheme (RAG). In RAG, interior nodes help to forward data for their sibling nodes and thus provide reliable data transmission for WSN. The simulation results demonstrate that AST can prolong the lifetime and RAG makes a better trade-off between storage and aggregation ratio, comparing to other aggregation schemes.

In [30], the authors proposed a novel Hierarchical Data Aggregation method using Compressive Sensing (HDACS), which combines a hierarchical network configuration with compressive sensing (CS). The authors’ main idea is to set multiple compression thresholds adaptively based on cluster sizes at different levels of the data aggregation tree to optimize the amount of packet delivery rate (PDR). The advantages of the proposed model in terms of the total amount of PDR and data compression ratio are analytically verified.

In [31], the authors presented the problem of scheduling virtual data aggregation trees to maximize the network lifetime when a fixed number of data are allowed to be aggregated into one packet, termed the Maximum Lifetime Data Aggregation Tree Scheduling (MLDATS) problem. The authors showed that the MLDATS problem is to be NP-complete. In addition, the authors proposed a local-tree-reconstruction-based scheduling algorithm (LTRBSA) for the MLDATS problem.

In [32], the authors applied data aggregation on two nodes levels in WSNs. They worked on sending fewer data from aggregator to the sink, along with the equation that expresses all data. They applied Bayesian belief network algorithm to measure the accuracy of the proposed scheme.

In [33], the authors propose a new scheme in the network layer, called Weighted Compressive Data Aggregation (WCDA), which benefits from the advantage of the sparse random measurement matrix to reduce the energy consumption. The novelty of the WCDA algorithm lies in the power control ability in sensor nodes to form energy efficient routing trees with focus on the load-balancing issue. In the second part, they present another new data aggregation method namely Cluster-based Weighted Compressive Data Aggregation (CWCDA) to make a significant reduction in the energy consumption in our WSN model. The main idea behind this algorithm is to apply the WCDA algorithm to each cluster in order to reduce significantly the number of involved sensor nodes during each CS measurement. In this case, candidate nodes related to each collector node are selected among the nodes inside one cluster. This yields in the formation of collection trees with a smaller structure than that of the WCDA algorithm

In this paper, a new method is proposed for WSNs which selects data aggregating nodes via COA. Indeed, the purpose of the study is to aggregate data with the aid of COA. Further, as our proposed method uses data aggregation spanning trees, it is only compared with other data aggregation spanning tree approaches such as PEDAPPA [19] and EESR [20] and GA [25]. In this paper, the frequency of usage of data aggregation tree phase is fixed and it is not dynamically adjusted.

3. The Proposed Method

In this section, we briefly describe our proposed scheme.

3.1 Preliminary Definitions

The primary model for WSNs which was presumed in this study is as follows:

- The network has *n* sensors which have been randomly and uniformly distributed in a pre-specified environment.
- Sensors and the main database have a fixed and certain location. They are not capable of moving.
- Nodes are distributed uniformly and randomly.
- The initial energy of the nodes is identical. It is equal to 1*J*.
- All the sensors can be detected and identified by means of their own unique IDs.

The following system model is used in this paper. The network consists of n wireless sensors, each containing a transceiver with a maximum transmission range. This work adopts the use of multi-hop transmissions in its communication model. For this purpose, a tree is constructed. In-network aggregation at intermediate nodes usually results in reducing the size of data that is forwarded by an intermediate node to its parent [7]. An undirected graph, i.e. G (V, E), was used for featuring the sensor network with V nodes which was produced based on the distance between nodes and the radio range. Random graph was produced in the following way: V sensors were randomly placed in a pre-specified environment where they were connected to their neighbors with the Euclidean distance equal to or less than their radio range.

Definition 1: in a data aggregation cycle, the data received from the environment is gathered by the intermediate nodes and transmitted to the base station.

Definition 2: data aggregation tree is a spanning tree with the root of base station which includes routing data for all the network nodes.

Definition 3: network lifetime is defined as the number of cycles of algorithm execution where all the network nodes are active.

Definition 4: the load of a sensor refers to the required energy for receiving and transmitting gathered data to its own father.

3.2 Cuckoo Optimization Algorithm (COA)

One notable search technique in computer science is to find the optimal solution in searching issues. COA is considered to be a supplementary algorithm which has been inspired from the life of a species of a bird known as cuckoo. It was developed in combination with levy flight instead of simple isotopic walk.

Evolution commences with a completely randomized set of entities and is repeated in the next generations. In each generation, the most appropriate ones but not the best ones are selected. A solution for a specific problem is illustrated through a list of parameters which are referred to as habitats. At the outset, several features are randomly produced for creating the first generation. During each generation, all the features are evaluated and the fitness value is measured by the fitness function. The next step is to produce the second generation of the community. For each individual, new locations are selected for the egg laying and migration stages. These selections are made in a way that they have the highest values with regard to the fitness function. Consequently, the most appropriate habitat with the highest benefit should be selected. The next steps are related to finding new locations with the maximum optimization benefit.

Since metaheuristic algorithm has no information about the global optimization point and the degree of the optimality of the solutions, certain criteria are needed for stopping them. The algorithm proposed in this study was designed in a way that it stops after the production of a number of generations. That is to say, if the populations of cuckoos get close to one another, practically, there will be no homogeneity and all the cuckoos reach an optimal point. In this case, algorithm will stop.

3.3 The Proposed Algorithm

Since the majority of data aggregation trees are produced based on remaining energy or the distance between sensors, hence, both parameters, namely the remaining energy and the distance between sensors were taken into consideration. As mentioned before, the algorithm proposed in this study for producing data aggregation tree is COA.

A. Habitat structure

Habitats provide data aggregation trees. Each habitat has a specific length which is appropriate for a specific number of available nodes in the network. The index of the array indicates the ID (identification) and the content of the array indicates the identity of its father. Figure 1 shows a habitat where the network includes 100 nodes. ID of the first node is 1 and the ID of its father is 5. Hence, the last node's father is node 43. The node with the zero value was regarded as the workstation or the root of the tree.

1	2	3	4	…	98	99	100
5	6	21	0	…	2	54	43

Fig. 1. Habitat structure

The production of the initial population and the next generations from it with regard to the above-mentioned structure of the habitat will be based on the following conditions:

- Each habitat should have one zero value (indicating workstation).
- The content of the array should not be the same as its ID.
- The content of the array has a random value between zero and *N* (*N* refers to the number of nodes).

Since the features within the habitats are randomly produced in creating the initial population, a node will be selected as the father where there is no mane between them in the graph. Thus, the produced habitat is invalid and other random numbers are produced.

B. Population

Population refers to a set of habitats. According to another definition, population includes valid data aggregation trees. For the initial population, parent nodes are randomly selected by acknowledging their validities.

C. Generation

The new generation is selected by applying egg laying operation and migration. The best habitat is selected based on the fitness function so that it is transmitted to the next generation.

1. Egg laying

In COA, the new generation is selected after the egg laying operation. Then, in the next stage, cuckoos move towards the optimal response with a little change. For using the positive characteristics of habitats, the operators are used in this way: each cuckoo randomly lays some eggs in the nest of the host bird which is located in the ELR (egg laying range) distance from it. Then, the profit function is determined for all the cuckoos. The best cuckoos with appropriate positions are selected according to the maximum number of cuckoos and are transmitted to the next generation. The remaining cuckoos which have less profit are destroyed. ELR (maximum egg laying range) is specified via equation (1):

$$\mathrm{ELR} = \alpha \times \frac{\text{Number of currcnt cuckoo}'\text{seggs}}{\text{Total number of eggs}} \times (\mathrm{var_{hi}} - \mathrm{var_{low}}) \quad (1)$$

In this equation, $\mathrm{var_{hi}}$ refers to the maximum number of each cuckoo's eggs, $\mathrm{var_{low}}$ denotes the minimum number of each cuckoo's eggs and α is a variable which adjusts the maximum value of ELR.

2. Migration

If only the egg laying operator is used for benefiting from the good features of habitats, the problem of premature convergence might occur. Hence, for preventing this problem and using the entire space for search, we use the migration operator. That is, when cuckoo chicks grow up and become mature, they live in the environment and their own groups for a while. However, when the egg laying time is approaching, they migrate to better habitats where there is more chance of being alive. After cuckoo groups in different areas are established, the average profit is measured so that the relative optimization of each group is obtained. For selecting the target point inside the chosen cluster with more average profit, the profit of all the cuckoos inside this cluster is measured and the cuckoo with the highest profit is selected as the target point. For specifying the new position of cuckoos around this point, the current position of the cuckoos is multiplied by the movement coefficient of the cuckoos and the new position of the cuckoos inside the optimal cluster is determined. The new position of the cuckoos after the migration operation is defined through Equation (2):

$$\mathrm{X_{Next\ habitat}} = \mathrm{X_{Current\ habitat}} + \mathrm{F} \times (\mathrm{X_{Goal\ point}} - \mathrm{X_{current\ habitat}}) \quad (2)$$

In Equation 2, F is a parameter which leads to deviation.

3. Correction function

Since different problems such as father-child, child-father problems, habitat without workstation and circle production might occur in the mentioned habitat, a correction function will be used for detecting and solving these problems which will prevent the transmission of invalid habitats to the next generation.

Habitat without work station: in habitat, the workstation has the value of zero. However, it is likely that no habitat has the value of zero for its content since the numbers are selected randomly. The following figure illustrates this problem for a network with 5 nodes.

1	2	3	4	5	6
3	2	4	2	3	5

Fig. 2. Habitat-related problem without workstation

Circle problem: this condition occurs when the content of array is the same as its index. Figure 3 shows this problem which occurs in the habitat.

1	2	3	4	5	6
3	1	6	4	3	5

Fig. 3. Circle problem

Father-child and child father problems: these problems occur when child a belongs to father b and when child *b* belongs to father *a* which is depicted in the following figure.

1	2	3	4	5	6
3	4	3	2	3	5

Fig. 4. Father-child, child-father problem

In case the circle problem occurs, one unit is added to the content of the array index. For solving the problem of residence location without workstation, zero value is randomly given to one of the elements of the array.

For sorting out the father-child, child-father problems, we begin with the first index which has a value other than zero. That is, we search for the related index in the elements of the array; if the obtained value has an index equal to the previous value, the problem has occurred; otherwise, we continue the procedure so as to navigate all the elements of the array. In case this problem occurs, one unit is added to the content of the related element. Figure 5 illustrates a pseudocode of the correction function which is executed for each habitat and it checks that the problems mentioned in the previous parts do not occur. If one of these problems occurs, the correction function will detect and correct it.

```
Correction function pseudo code
Procedure Repair function (habitat)
For each habitat_i in habitat do
While habitat_i creates a cycle do
        Replace value of habitat_i with (value of habitat_{i+1})
While habitat have not a sink do
        Randomly insert zero In the habitat_i
While habitat have a problem Father-Child and Child-Father do
    Replace value of habitat_i with (value of habitat_{i+1})
While habitat_i is not in range of board do
        Replace value of habitat_i with other node   ID
End.
```

Fig. 5.Correction function pseudo code

4. Evaluation and fitness

The algorithm proposed in this study has two purposes. Firstly, the tree should be effective with regard to energy consumption so that nodes can have more communications for long-term frequencies. Secondly, the tree produced among the nodes should be balanced. The fitness of the habitats is specified through the following parameters:

Energy expenditure rate: a sensor has different expenditures in producing different trees. The tree produced by the COA algorithm needs effective energy. It is preferred that each node consumes little energy in each communication cycle. For each *i* node, the energy consumption rate is measured through equation (3). This rate increases for all the sensor nodes.

$$\mathrm{L} = \frac{\mathrm{e_i}}{\mathrm{E_{rx}}} \tag{3}$$

In this equation, e_i stands for the current energy of each node, E_{rx} denotes the amount of energy consumption for receiving data packets from the child. E_{rx} is determined through equation (4).

$$E_{rx} = k \times E_{elec} \tag{4}$$

In this equation, E_{elec} denotes the required energy for maintaining movement and traffic within the sensors and *K* stands for the number of available bits in the packet.

In this paper, the following generic model is used to account for the energy consumed during communication:

$$E_{Tx}(i,j) = k[E_{elec} + \varepsilon_{amp} \times d^2] \tag{6}$$

$$E_{Total}(i,j) = k[2E_{elec} + \varepsilon_{amp} \times d^2] \tag{7}$$

In Equations (6) and (7), *d* is Euclidian distance between two neighbor nodes *i* and *j*, ε_{amp} is he amplification energy required to transmit a bit a unit distance and $E_{Total}(i,j)$ is the total energy consumption for transmission and receiving *k* bits of data from node *i* to node *j*. For sake of simplicity, we assume that the energy consumed in the sleep state is negligible.

Method of measuring e_i: e_i refers to the degree of current energy. The initial value of e_i is considered to be 1*J* but this value varies during the execution of the algorithm so that after the execution of each cycle, the new $\mathrm{e_i}$ is specified via Equation 8.

$$r_i = e_i \text{-} E_{rx} \tag{8}$$

In this equation, r_i stands for the remaining energy of the *i*th node after the execution of a cycle. That is, the remaining energy of each node at the end of a cycle is considered to be the current energy of it in the next cycle.

Standard deviation of the remaining energy: after each cycle, the amount of energy E_{rx} will be consumed by the i sensor (n_i). For maintaining the nodes in the working condition, it is essential that the energy expenditures be distributed in a balanced way based on the remaining energy of the sensors. Standard deviation of the sensors of the data aggregation tree is regarded as an appropriate criterion for expressing energy expenditure among nodes. The lower STD (r_i), the higher the network lifetime. The parameter of standard deviation is measured through equation (9):

$$\sigma = \sqrt{{}^{1}\!/_{N} \sum_{i=1}^{N} (e_i - e^-)^2} \tag{9}$$

In equation (6), e^- is determined via equation (10):

$$e^- = {}^{1}/_{N} \sum_{i=1}^{N} e_i \qquad (10)$$

The distance between the sesonrs: the distance between the sensors is one of the parameters which can be taken into consideration for producing data aggregation tree since the closer is the father node to the children, the lower will be the cost of transmitting and receiving data.

$$\mathrm{d}\ (parent, child\ ) = \sqrt{(x_p - x_{ch})^2 + \ (y_p - y_{ch})^2} \qquad (11)$$

In this Euclidean equation, $(\boldsymbol{x_p}, \boldsymbol{y_p})$ denote father's location coordinates and $(\boldsymbol{x_{ch}}, \boldsymbol{y_{ch}})$ stand for the child's location coordinates.

Energy on level rate: since data aggregation task in the tree is done by the father nodes, hence, the higher is the father's energy compared with child's energy, the better it will be. For examining this issue, *E* parameter is used which is determined through the following procedure:

1. *E*=0 (the value of this parameter is initially zero)
2. For the correctness of the Equation ${}^{e_p}/_{e_{ch}} > 1$, one unit is added to *E* in one element of the array (*E=E+1*).

Data aggregating nodes are selected based on the following parameters: nodes' energy level, energy expenditure rate, distance between father and children nodes, and standard deviation of the remaining energy of the nodes. Based on these parameters, a degree is determined for each node and the node with the highest degree of fitness will be selected as the data aggregator. This grading is conducted using the following Equation:

$$F_i \ = \alpha_1(E) + \alpha_2(\tfrac{1}{D}) + \ \alpha_3(L_i) + \alpha_4(\tfrac{1}{std}) \qquad (12)$$

In this equation, $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_{2,}}, \boldsymbol{\alpha_3}\ and\ \boldsymbol{\alpha_4}$ denote the weight of the parameters and the values are considered in a way that they fit the units of the parameters and would be true in Equation (13).

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1 \qquad (13)$$

Since of the parameters values of the equations were simulated for several times, the obtained values for the above-mentioned parameters are as follows:

$\alpha_1 = 0.25$,$\alpha_2 = 0.25$ m, α_3=0.25, $\alpha_4 = 0.25$ j

Each habitat is measured based on target function so that the habitat with the highest profit is transmitted to the next generation.

4. Performance Evaluation

This section shows the effect of the proposed algorithm on network lifetime and energy consumption. In the following analysis, the sensor nodes are considered to be uniformly and randomly placed in a two-dimensional area. All sensor nodes are homogeneous and have equal, omnidirectional transmission patterns of range and the sink energy supply is adequate. The efficiency of the proposed algorithm was investigated via different tests and simulations which were conducted in MATLAB R2009a. It should be noted that the performance and the efficiency of the proposed algorithm was compared with those of GA [25], PEDAPPA [19] and EESR [20]. In this paper and in simulation phase, we assume that all nodes perform in-network data aggregation and all the sensor nodes in the network are static and are loosely synchronized to enable TDMA (time division multiple access) scheduling. The topology graph constructed from the network is connected. Weights associated with all the edges are positive and hence, there are no negative weight cycles. A node can adjust its power level depending on its requirements. The results for lifetime are obtained for the 95% confidence interval.

The comparison of these algorithms were done in different dimensions of the network with respect to network lifetime. The simulation parameters used in different tests and analyses are given in table 1.

Table 1. Parameters used in the simulation

Parameters	Values
Initial energy	1J
Number of sensors	100
Packet size	1000 Byte
Radio range	25 m
Cuckoos' movement coefficient	2
Maximum egg laying radius	15
E_{amp}	100 pJ/bit/m^2
E_{elec}	50 nJ/bit

In the conducted experiments, simulations were proceeded until the first node was destroyed. Indeed, the simulations were carried out in 10 rounds and each experiment was executed for three times and the average of the three executions were taken into consideration. The location of the base station in all the experiments was considered to be at the center of the sensing environment. Figures 6, 7 and 8 depict network lifetime for the network sizes of 50m*50m, 100m*100m and 200m*200m, respectively. The results of the simulations indicated that the proposed algorithm had better performance than GA [25], PEDAPPA [19] and EESR [20].

We have compared network lifetimes obtained in different networks with different number of sensors. As can be seen in the Figures, our proposed COA notably increases network lifetime.

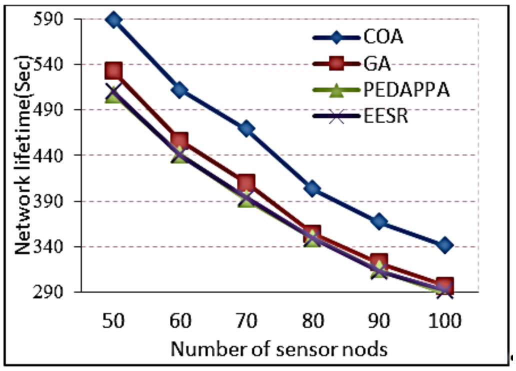


Fig. 6. Network lifetime in sensing environment with the size of 50m*50m

GA [25], PEDAPPA [19] or EESR [20] does not consider the maximal load of all sensors, while COA does. Therefore, network lifetimes obtained by GA [25], PEDAPPA [19] and EESR [20] are shorter than network lifetimes obtained by COA which reduces the maximal load of all sensors.

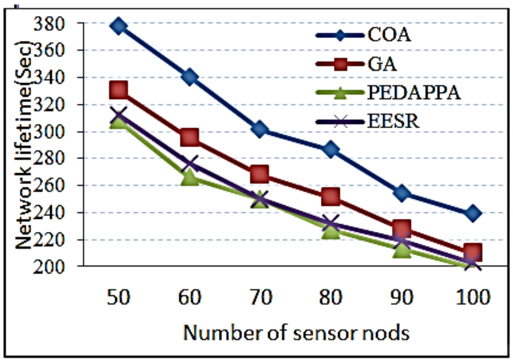


Fig. 7. Network lifetime in sensing environment with the size of 100m*100m

In PEDAP-PA [19], the sink needs to broadcast the topology of the network to all nodes in the network periodically; hence it consumes most energy. Due to fitness function of COA, it improves network lifetime in comparing with GA [25], PEDAPPA [19] and EESR [20].

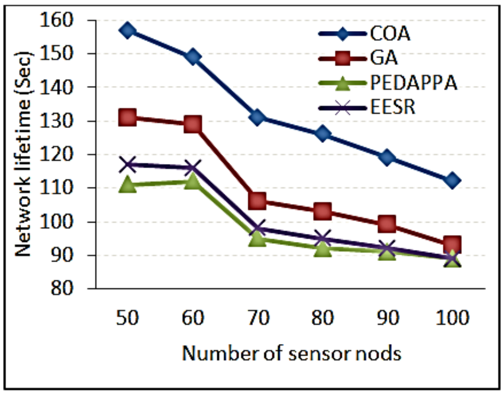


Fig. 8. Network lifetime in sensing environment with the size of 200m*200m

Figure 9 shows the remaining energy of networks after 100 rounds execution for CoA, PSO [34] (Particle Swarm Optimization) and ICA [35] (Imperial Competitive Algorithm). As Figure 9 depicts, the reaming energy for proposed scheme (COA) is better than PSO [34] and ICA [35]. Hence, the proposed scheme is energy efficient and improves network lifetime.

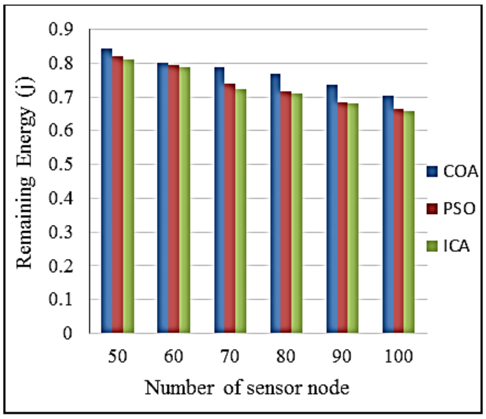


Fig. 9. Remaining energy of network after 100 rounds execution

Figure 10 shows the convergence diagram for proposed scheme.

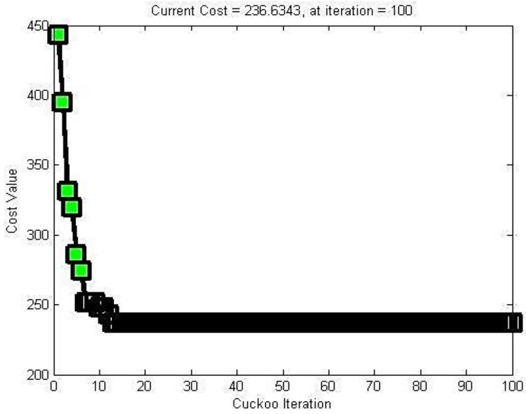


Fig. 10. Convergence diagram for proposed scheme (COA)

5. Conclusions

In this paper, a new method based cuckoo algorithm was proposed for aggregating data in WSNs. In this method, data aggregation tree was produced based on the following parameters: energy expenditure rate, standard deviation of the sensors' remaining energy, energy on level rate and the distance between the sensors. The proposed scheme for data aggregation trees extend the network lifetime as compared to EESR, PEDAPPA and GA. It was found that the proposed algorithm can reduce the energy consumption of the network significantly; hence, network lifetime is consequently enhanced. To sum it up, it can be concluded that the proposed method has desirable performance and efficiency. As a future work, we would like to investigate adaptive tree structure frequency techniques and we would like to use the proposed scheme in WSNs based on software defined networks (SDNs) [36] technology.

References

[1] J. Yick, B. Mukherjee, and D. Ghosal. “Wireless sensor network survey.” Computer networks, vol. 52, pp. 2292-2330, 2008.

[2] A. Ghaffari. “Congestion control mechanisms in wireless sensor networks: A survey. ” Journal of Network and Computer Applications, vol. 52, pp. 101-115, 2015.

[3] A. Ghaffari. “An energy efficient routing protocol for wireless sensor networks using A-star algorithm.” Journal of applied research and technology, vol. 12, pp. 815-822, 2014.

[4] Z. Mottaghinia and A. Ghaffari. “A Unicast Tree-Based Data Gathering Protocol for Delay Tolerant Mobile Sensor Networks.” Information Systems & Telecommunication, pp. 59, 2016.

[5] X. Zhu, G. Chen, S. Tang, X. Wu, and B. Chen. “Fast Approximation Algorithm for Maximum Lifetime Aggregation Trees in Wireless Sensor Networks.” INFORMS Journal on Computing, vol. 28, pp. 417-431, 2016.

[6] A. Ghaffari, L. Darougaran, and A. Shiri. “Comparing data aggregation methods in wireless sensor networks.” presented at the Third national symposium on computer engineering and information technology, 2010.

[7] S. Upadhyayula and S. K. Gupta. “ Spanning tree based algorithms for low latency and energy efficient data aggregation enhanced convergecast (dac) in wireless sensor networks. Ad Hoc Networks, vol. 5, pp. 626-648, 2007.

[8] A. jhumka, M. Bradbury, and S. Saginbeko. “Efficient fault-tolerant collision-free data aggregation scheduling for wireless sensor networks.” Parallel Distributed Computing, vol. 74, pp. 1789-1801, 2014.

[9] R. R. Rout and S. Ghosh. “Adaptive data aggregation and energy efficiency using network coding in a clustered wireless sensor network: An analytical approach.” Computer Communications pp. 65-75, 2014.

[10] SH. Niu, C. Wang, Z. Yu, and S. Cao. “Lossy data aggregation integrity scheme in wireless sensor networks.” Computers and Electrical Engineering, vol. 39, pp. 1726-1735, 2013.

[11] S. park. “ performance Analysis of Data Aggregation Schemes For wirless sensor network, 2006

[12] S. Sicari, L. Grieco, G. Boggia, and A. Coen-Porisini. “DyDAP: A dynamic data aggregation scheme for privacy aware wireless sensor networks.” The Journal of Systems and Software, vol. 85, pp. 152-166, 2012.

[13] Q. Liua, Y.Changa, and X. Jiab. “A hybrid metod of CSM/CA and TDMA for real-time data aggregation in wireless sensor networks.” Computer Communications, vol. 39, pp. 269-278, 2013.

[14] W. Wu, J. Cao, H. Wu, and J.Li. “Robust and dynamic data aggregation in wireless sensor networks: A cross-layer approach.” Computer Networks, vol. 57, pp. 3929-3940, 2013.

[15] H. Yousefi, M. Yeganeh, N.Alinaghipour, and A. Movaghar. “Structure-free real-time data aggregation in wireless sensor networks.” Computer Communications, vol. 35, pp. 1132-1140, 2012.

[16] H.Li, CH.Wua, Q. Huab, and F. Lau. “Latency-minimizing data aggregation in wireless sensor networks under physical interference mode.” Ad Hoc Networks, vol. 12, pp. 52-68, 2014.

[17] S. Ozdemir and Y.Xiao. “Integrity protecting hierarchical concealed data aggregation for wireless sensor networks.” Computer Networks, vol. 55, pp. 1735-1746, 2011.

[18] R. Rajabioun. “Cuckoo Optimization Algorithm.” Applied soft computing, vol. 11, pp. 5508- 5518, 2011.

[19] H. Ö. Tan and I. Körpeoğ lu. “Power efficient data gathering and aggregation in wireless sensor networks.” ACM Sigmod Record, vol. 32, pp. 66-71, 2003.

[20] S. Hussain and O. Islam. “An energy efficient spanning tree based multi-hop routing in wireless sensor networks.” in 2007 IEEE Wireless Communications and Networking Conference, 2007, pp. 4383-4388.

[21] P.Wang, Y. He, and L.Huang. “Near optimal scheduling of data aggregation in wireless sensor networks.” Ad Hoc Networks, vol. 11, pp. 1287- 1296, 2013.

[22] CH.Chaon and T. Hsiao. “Design of structure-free and energy-balanced data aggregation in wireless sensor networks.” vol. 37, pp. 229-239, 2014.

[23] H.Liu, Z.Liu, D.Li, X.Lub, and H.Due. “Approximation algorithms for minimum latency data aggregation in wireless sensor networks with directional antenna.” Theoretical Computer Science, vol. 497, pp. 139-153, 2013.

[24] W. -H. Liao, Y. Kao, and C.-M. Fan. “An ant colony algorithm for data aggregation in wireless sensor networks.” in Sensor Technologies and Applications, 2007. SensorComm 2007. International Conference on, 2007, pp. 101-106.

[25] O. Islam, S. Hussain, and H. Zhang. “Genetic Algorithm for Data Aggregation Trees in Wireless Sensor Networks.” 2007.

[26] S. Lai and B.Ravindran. “Achieving Max–Min lifetime and fairness with rate allocation for data aggregation in sensor networks.” Ad Hoc Networks, vol. 9, pp. 821-834, 2011.

[27] S. Kwond, J. H. Ko, and J.Kim. “Dynamic timeout for data aggregation in wireless sensor networks.” Computer Network, vol. 55, pp. 650-664, 2011.

[28] J.AL-Karaki, R.Mustafa, and A.Kamal. “Data aggregation and routing in Wireless Sensor Networks: Optimal and heuristic algorithms.” Computer Network, vol. 53, pp. 954-960, 2009.

[29] Y. Zhang, J. Pu, X. Liu, and Z. Chen. “An adaptive spanning tree-based data collection scheme in wireless sensor networks.” International Journal of Distributed Sensor Networks, vol. 2015, p. 2, 2015.

[30] X. Xu, R. Ansari, A. Khokhar, and A. V. Vasilakos. “Hierarchical data aggregation using compressive sensing (HDACS) in WSNs.” ACM Transactions on Sensor Networks (TOSN), vol. 11, p. 45, 2015.

[31] N.-T. Nguyen, B.-H. Liu, V.-T. Pham, and Y.-S. Luo. “On maximizing the lifetime for data aggregation in wireless sensor networks using virtual data aggregation trees.” Computer Networks, vol. 105, pp. 99-110, 2016.

[32] I. Atoui, A. Ahmad, M. Medlej, A. Makhoul, S. Tawbe, and A. Hijazi. “Tree-based data aggregation approach in wireless sensor network using fitting functions.” in 2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC), 2016, pp. 146-150.

[33] S. Abbasi-Daresari and J. Abouei. “Toward cluster-based weighted compressive data aggregation in wireless sensor networks.” Ad Hoc Networks, vol. 36, pp. 368-385, 2016.

[34] J. Kennedy and R. Eberhart. “Particle swarm optimization.” in Neural Networks, 1995. Proceedings., IEEE International Conference on, 1995, pp. 1942-1948 vol.4.

[35] E. Atashpaz-Gargari and C. Lucas. “Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition.” in 2007 IEEE Congress on Evolutionary Computation, 2007, pp. 4661-4667.

[36] R. Masoudi and A. Ghaffari. “Software defined networks: A survey.” Journal of Network and Computer Applications, vol. 67, pp. 1-25, 2016.

Elham Mohsenifard is a M.Sc graduate of Computer Engineering at Tabriz branch, Islamic Azad University, Tabriz, Iran. She received her B.Sc degree from Islamic Azad University (IAU). Her research interests include Wireless Sensor Networks (WSNs) and Mobile Ad Hoc Networks (MANETs).

Ali Ghaffari received his B.Sc, M.Sc and Ph.D. degrees in Computer Engineering from the University of Tehran and IAUT (Islamic Azad University), TEHRAN, IRAN in 1994, 2002 and 2010 respectively. As an assistant professor of computer engineering at Islamic Azad University, Tabriz branch, IRAN, his research interests are mainly in the field of wired and wireless networks, Wireless Sensor Networks (WSNs), Mobile Ad Hoc Networks(MANETs), Vehicular Ad Hoc Networks(VANETs), networks security and Quality of Service (QoS). He has published more than 60 international conference and reviewed journal papers.

Instance Based Sparse Classifier Fusion for Speaker Verification

Mohammad Hasheminejad
Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran
mhashemi@birjand.ac.ir
Hassan Farsi*
Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran
hfarsi@birjand.ac.ir

Received: 28/Feb/2016 Revised: 26/Apr/2016 Accepted: 07/May/2016

Abstract

This paper focuses on the problem of ensemble classification for text-independent speaker verification. Ensemble classification is an efficient method to improve the performance of the classification system. This method gains the advantage of a set of expert classifiers. A speaker verification system gets an input utterance and an identity claim, then verifies the claim in terms of a matching score. This score determines the resemblance of the input utterance and pre-enrolled target speakers. Since there is a variety of information in a speech signal, state-of-the-art speaker verification systems use a set of complementary classifiers to provide a reliable decision about the verification. Such a system receives some scores as input and takes a binary decision: accept or reject the claimed identity. Most of the recent studies on the classifier fusion for speaker verification used a weighted linear combination of the base classifiers. The corresponding weights are estimated using logistic regression. Additional researches have been performed on ensemble classification by adding different regularization terms to the logistic regression formulae. However, there are missing points in this type of ensemble classification, which are the correlation of the base classifiers and the superiority of some base classifiers for each test instance. We address both problems, by an instance based classifier ensemble selection and weight determination method. Our extensive studies on NIST 2004 speaker recognition evaluation (SRE) corpus in terms of EER, minDCF and minCLLR show the effectiveness of the proposed method.

Keywords: Speaker Recognition; Speaker Verification; Ensemble Classification; Classifier Fusion; IBSparse**.**

1. Introduction

Scientific studies have shown that, there are varieties of information in a speech signal which can help speaker recognition. Speaker recognition is a process of decision making about a speaker's identity using the person's speech signal. The field of speaker recognition contains two main branches; speaker verification and speaker identification. In speaker verification, an identity claim is first constructed and then the claim is accepted, or rejected, based on the information extracted from the corresponding speech signal. On the other hand, a speaker identification system, at first, registers a set of target speakers and then determines the identity of the owner of an incoming speech signal. Since a speaker verification system can lead to speaker identification and there are more sophisticated criteria to evaluate a speaker verification system, the majority of speaker recognition research is devoted to speaker verification tasks. To gain advantage of different information of speech in the verification process, an ensemble of base classifiers can be used. Classifier fusion is an important subject in speaker verification which can be performed on the feature, score or decision level [1]. On the feature level fusion, different feature vectors are concatenated to construct a new feature vector. In speaker verification, fusion of scores includes obtaining matching scores for each base classifier and obtaining a final score from these base scores using a proper role. On the decision level fusion, the final decision is a logical fusion of decision output of different classifiers or modalities. This logical fusion can be *"AND"*, *"OR"* or a combination of both. In this paper, we focus on score level fusion where the final score is a weighted summation of base scores. Contrary to most of the classifier fusion for speaker verification works, which use the weighted sum of scores for all test instances [2], or a simple arithmetic mean of scores as the final score [3], we use an instance-specific ensemble of classifiers whose weights are adopted separately for each test instance. Despite that using permanent weights for score fusion in a speaker verification task, may be effective in some situations, obtaining optimum weights which are effective for all test instances is troublesome. In these methods a set of unique weights are learned on training or held back data. Thereafter, these weights are used in the verification process of all trials. This may not be generalizable to all test samples, since some base classifiers may be effective for some of the test samples and not for others. In this paper, we consider this issue and exploit the instance-specific behavior speaker classifiers. To do this we were inspired by [2], [4] and [5], and act in the following procedures:

* Corresponding Author

- We determine the weight of each classifier according to the test instance. Then, we calculate the final score as a weighted sum of scores obtained from all base classifiers.
- We also consider sparse classifier fusion using the behavior. Therefore, in this case, the final score is a weighted sum of scores from a few base classifiers.
- We introduce a new formula to determine the final fusion score.

Logistic regression with elastic-net regularization is also considered as a baseline to show the effectiveness and generalization power of the proposed method.

2. Base Classifiers

A typical speaker verification system consists of train and test phases. In the train phase we introduce target speakers to the system. In the test phase incoming unlabeled utterances are claimed as belonging to one of the enrolled targets and the system verifies validity of this claim. Figure 1 shows workflow of such a system. Speaker feature extraction methods transform the original speech signal to a compact representation. These methods aim at holding speaker specific information it the resulting representation.

To create powerful base classifiers, we used four widely used speech features in speaker recognition. These features are mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), stabilized weighted linear prediction (SWLP) [6], and linear predictive cepstral coefficient (PLCC) [7]. Conventional linear prediction (LP) determines a pth order autoregressive model for a speech frame, by minimizing the sum of squares of prediction errors. Weighted linear prediction (WLP) is obtained by introducing temporal weighting of the squared prediction error. The SWLP which is used in our research, is a variant of WLP that guarantees the stability of WLP filter. A concise description of MFCC and PLP feature extraction is provided in [8].

In the matching step, the system tries to specify the similarity of enrolled target features (templates) and incoming speech features, in terms of a verification score. In the late steps of the verification, the score is compared to a predefined threshold value. This threshold value is computed from training data. If the score of the trial is more than the threshold, the claim is accepted, otherwise it is rejected.

We used three different powerful modeling methods of speakers. These methods are, GMM-SVM-KL [9], GMM-SVM-BHAT [10] and ivector-PLAD [11]. SVM based methods have been successful in text independent speaker verification. In these methods, the SVM is combined with the GMM supervector concept. They derive a kernel (we used Kullback-Leibler (KL) divergence and Bhattacharya (BHAT) distance), then apply the Nuisance Attribute Projection (NAP) [12] to the kernel. Total variability or ivector systems provide an elegant way of dimensionality reduction of speech features. This technique converts a sequence of feature frames to a fixed length low dimensional vector. This vector represents the whole utterance (i.e. the whole speaker) and can be an input to a standard pattern recognition algorithm. We then use Probabilistic Linear Discriminant Analysis (PLDA) for scoring.

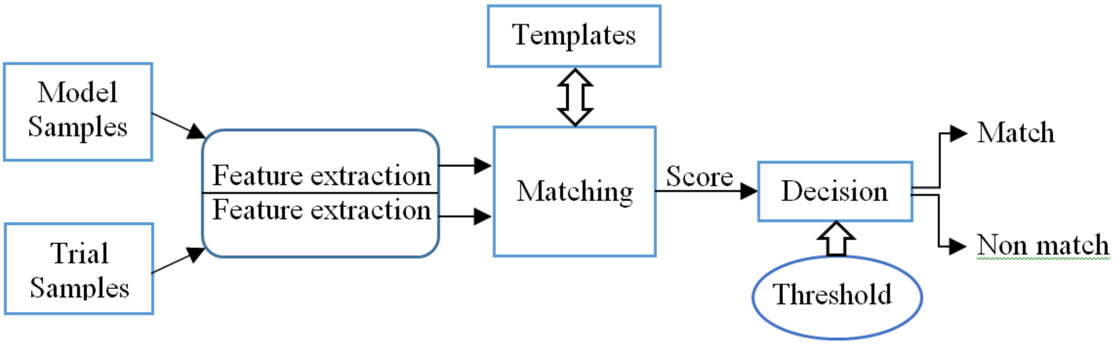


Fig. 1. Block diagram of a typical speaker verification system

3. Score Fusion in Speaker Verification

There are three main levels on which biometric classifier fusion can occur: feature level fusion, score level fusion and decision level fusion. Feature level fusion (early fusion methods) occurs before the invocation of the matching block (Figure 1). In this process a new feature vector is created. This new feature vector is a combination of previously extracted features. The matching step is performed on the new feature vector. In score level fusion (late fusion methods), which is our main subject, some basis expert classifiers are firstly employed to obtain the matching score between each test sample, and the previously stored templates. It is shown that score level fusion methods provide better results than feature level fusion [5]. In decision level fusion approaches, accept or reject decision of individual classifiers serve as input to the fusing function [13].

In the case of score fusion, the final score is obtained as, $s_f = w_0 + \sum_{l=1}^{L} w_l s_l$, in which L is number of base classifiers, w_0 is added to calibrate the final score and w_l and s_l are weight and score of l^{th} classifier, respectively. Regardless of training individual classifiers, there is a need to train the fusion process. An intuitive way of obtaining a final score from ensemble classifier scores is to estimate the fixed weight for each classifier. To do this one needs a set of labeled scores, whose labels are either 0 ($y_i = 0$) if the utterance belongs to the claimed target or 1 ($y_i = 1$) if the utterance does not belong to the claimed target. If the number of training scores is N_{dev} , a development set $D = \{(s_i, y_i, I = 1, \dots, N_{dev})\}$ is used to train this model. Such a system does not need to know anything about how individual classifiers are trained or the speech features. After selecting proper training scores an optimization method is employed to minimize an error criterion or maximize an efficiency measure. The optimization can be directly performed using a neural network [14], heuristic algorithms [15] or the widely used logistic regression [2].

3.1 Logistic Regression Based Fusion

State-of-the-art speaker verification systems use multiple classifiers to make a reliable decision. Linear regression is a discriminative model [16] which is commonly used to fuse scores in speaker verification. In this section we explain why this method is widely used and accepted in speaker verification and how it is improved in recent years.

In test phase of an ensemble speaker verification system, an identity is firstly claimed for an incoming utterance signal. Each classifier in the ensemble, measures validity of the claim, in terms of similarity score. At this stage the system needs a score fusion method to accept or reject the claim. The score fusion method should realize a mapping from $\mathbb{R}^n$ space to a binary space {0, 1}, where 0 means the identity claim is accepted and 1 means it is rejected. We can cast the problem as two target classification problems with an n dimensional input feature vector. Elements of these vectors should be of the same type, e.g. probability. One may calculate best weights, which minimize classification error for the training set, using a brute force approach. But, there is a question of generalization. There is considerable variation between training and runtime scores. This is why it is recommended to use estimates of real probabilities as scores [17]. Bayesian framework [16], which minimizes classification error probability, can be used to reach those probabilities. We will provide a general overview of the Bayesian decision rule next here.

Suppose there are two classes, T and I, representing target and non-target (Imposter) classes, respectively. For a given random score vector of X, which may belong to either of class the cost of classifying a class i score vector into a class j event, can be a zero-one loss function (Equation (1)):

$$\mathrm{C_{ij}} = \begin{cases} 0 & \text{if } \mathrm{i = j} \\ 1 & \text{if } \mathrm{i \neq j} \end{cases} \tag{1}$$

This assigns zero loss to a correct classification and a unit loss to a miss classification. Under this assumption, Bayes rule defines the posterior probability of class i as Equation (2):

$$\mathrm{P(c_i|X) = \frac{P(X|c_i).P(c_i)}{P(X)}} \tag{1}$$

Where $P(X)$ is prior probability of X and $P(c_i)$ is prior probability of c_i. We need to estimate probability distribution correctly, to get reliable scores. In [17] it is explained how we can reduce Equation (2) to $P(X|c_i)$ using assumptions about prior probabilities. If we suppose different base classifiers are independent of one another, $P(X|c_i)$ results in equation (3):

$$\mathrm{P(X|c_i) = P(s_1, \dots, s_K|c_i) = \prod_{k=1}^{K} P(s_k|c_i)} \tag{2}$$

Where K is the number of base classifiers, c_i is a label for class i and s_k is the k^{th} element of score vector. Since the scores for T class (target) are correlated, the recent assumption is not intuitive. This is due to the fact that if a trial belongs to a target class, scores of all good classifiers are close to unity. It is more reasonable to believe that s_k for the imposter and $(1 - s_k)$ for the target are not correlated. The posterior probability of target class can be derived as equation (4) [17]:

$$\mathrm{P(T|s_1, \dots, s_k) = \frac{1}{1 + e^{-\{(\sum_{k=1}^{K} x_k) + x_0\}}}} \tag{3}$$

Where $\mathrm{x_0 = \ln\frac{P(T)}{P(I)}}$, and $\mathrm{x_k = \ln\frac{P(s_k|T)}{P(s_k|I)}}$.

If we suppose that the probabilities are members of the exponential family (equations (5) and (6)):

$$\mathrm{P(s_k|T) = f(s_k).\, e^{(C_k.s_k + C_{k0})}} \tag{4}$$

$$\mathrm{P(s_k|I) = f(s_k).\, e^{(C_k.s_k + C_{k0})}} \tag{5}$$

Then equation (4) is reduced to logistic regression (LR) model or logistic distribution function (equation (7)):

$$\mathrm{P(T|s_1, \dots, s_k) = \frac{1}{1 + e^{-g(s)}} = \pi} \tag{6}$$

Where:

$$\mathrm{g(s) = \beta_0 + \beta_1.s_1 + \cdots + \beta_K.s_K} \tag{7}$$

$$\mathrm{\beta_0 = \sum_{k=1}^{K} (C_{k0} - I_{k0}) + \ln\frac{P(T)}{P(I)}} \tag{8}$$

$$\mathrm{\beta_K = C_k + I_k} \tag{9}$$

A particular case of the exponential family is a Gaussian distribution. If we suppose distribution of the classes are Gaussian, equations (9) and (10) become equal to equations (11) and (12).

$$\beta_0 = \sum_{k=1}^{K} \frac{(\mu_k^I)^2 - (\mu_k^T)^2}{2\sigma_k^2} + \ln \frac{P(T)}{P(I)} \tag{10}$$

$$\beta_K = \frac{\mu_k^T - \mu_k^I}{\sigma_k^2} \tag{11}$$

Where μ_k^T and μ_k^I are the mean of the target and imposter distributions, respectively, and σ_k^2 is the common variance. An interesting result of this method is that, the weight of the kth classifier, β_K, is proportional to the difference of the means of, the target and imposter distributions and if the classifier has scattered target scores it is not reliable and has a lower weight.

To this point we should find optimal weights (β_K in equation (8)), so that $P(T|s_1, \dots, s_k)$, indicated by equation (7), is maximized. To solve the problem, researchers took many issues in to consideration and introduced cost functions. One of the most recent defined cost functions is $C_{wlr}(w, D)$[2] which is given by:

$$\begin{aligned} C_{wlr}(w) = \frac{P_{eff}}{N_t} \sum_{i=1}^{N_t} \log\left(1 + e^{-w^T s_i - \text{logit}\, P_{eff}}\right) \\ + \frac{1 - P_{eff}}{N_f} \sum_{j=1}^{N_f} \log\Big(1 \\ + e^{w^T s_j + \text{logit}\, P_{eff}}\Big) \end{aligned} \tag{12}$$

Where $P_{eff} = logit^{-1}\left(logit(P_{tar}) + log\left(C_{miss}/C_{fa}\right)\right)$, depends on the prior probability of a target speaker (P_{tar}), the cost of miss classification (C_{miss}) and the cost of false acceptation (C_{fa}). The aim of defining such a cost function is to find the optimal weights which minimize the cost function. Equation (14) formulates this optimization problem:

$$w^* = \underset{w}{\text{argmin}}\, C_{wlr}(w) \tag{13}$$

This formulation changed to (15) when V. Hautamäki, et.al. showed in [2] that, a regularized version of equation (14), which takes a sparse number of classifiers in the ensemble, acts better. This optimization problem is regularized using a combination of ridge and LASSO regressions which is called elastic-net.

$$\begin{aligned} w^* = \underset{w}{\text{argmin}}\{C_{wlr}(w) \\ + \lambda(\alpha\|w\|_1 + (1-\alpha)\|w\|_2^2)\} \end{aligned} \tag{14}$$

Where coefficient, λ, which is a Lagrange multiplier, determines the amount of shrinkage of the weights. The constraint $\|w\|_1$ is known as LASSO and $\|w\|_2^2$ corresponds to ridge regression.In elastic-net, the LASSO part causes most of the weights to be near zero. This means that, it is a sparsity promoting constraint. The other part of the elastic-net is ridge constraint, which causes the weights not to push as aggressively as LASSO only constraint. Coefficient α determines the amount of participation of the LASSO and ridge in the equation. This problem can be solved using the ProjectL1 algorithm [18][1].

Although this method tries to increase the generalization capability of the classifier fusion, it identifies a unique weight for every classifier. These weights are obtained from training or held-out data and are used on all instances. This method also chooses a sparse number of classifiers during the training process and omits most of them from the test process. We show that a classifier, which is omitted from the ensemble set, has better performance for specific test instances. Recent studies showed that, taking the instance based behavior of classifiers improves generalization of the ensemble classifier [4],[5]. In the following section we introduce the proposed method to take the instance base behavior of speaker verification experts.

3.2 Instance Based Ensemble and Weight Selection

Weights which are selected based on a test sample, should score the prediction capability of each classifier on that sample. If the test sample in a trial is *positive* (real target), and the score of a classifier is high, then its weight should be high, and the weight should be low when this score is low. If the sample is *negative* (is not target), and the score of a classifier is high its weight should be low and if its score is low, it means that the classifier has made a reasonable decision, and the weight should be high. We call the proposed method instance based sparse classifier fusion (IBSparse).

Discovering individual weights for each trial is a challenging task. Since there is no information about the real label of the trial, we do not know if the classifier decision is correct or not, and as a result, it is not clear how to derive a specific weight for the classifier.

3.2.1 Clarity Index

Clarity index is an objective that can be used to obtain sample specific weights [4]. This objective is based on test scores and previously obtained training scores and has nothing to do with low level features. Each classifier has n_0 positive and n_1 negative training scores, which are obtained in the training phase. The positive scores are those scores whose related utterance originated from the target and negative scores are scores that belong to the impostor utterances. The Clarity index depends on two factors. The first factor is Relevance Loss (RL) which determines the position of a test score vector, S^{ts}, against negative training scores (S_i^{ntr}). Equation (16) defines RL:

$$RL(S^{ts}, W) = \frac{1}{n_n} \sum_{i=1}^{n_n} U(W^T S_i^{ntr} - W^T S^{ts}) \tag{15}$$

Where W is the weight vector, S^{ts} is the test score vector, n_n is the number of negative training scores and U is the unit step function. RL is a fraction of the non-target

[1] Available online at: "http://www.cs.ubc.ca/~schmidtm/Software/code.html"

training scores divided by the total number of non-target scores. Therefore, this value is in the range of [0,1]. For a target trial, the ideal state is that the test score will be higher than all negative training scores. As a result, this value is desired to be close to 0. For a non-target trial, the ideal state is that the test score would be less than all negative training scores, consequently, the value is desired to be close to one.

The second factor is irrelevance loss (IL) which determines the position of a test score vector against the positive training scores (S_i^{ptr}). Equation (17) defines IL:

$$\mathrm{IL}(\mathrm{S^{ts}},\mathrm{W}) = \frac{1}{\mathrm{n_p}}\sum_{\mathrm{j=1}}^{\mathrm{n_p}} \mathrm{U}\left(\mathrm{W^T S^{ts}} - \mathrm{W^T S_j^{ptr}}\right) \qquad (16)$$

Where n_p is the number of positive training scores. IL is desired to be close to 1 for a target trial and 0 for a non-target trial.

The raw clarity index is then defined as the difference between RL and IL (Equation (18)):

$$\mathrm{RCL}(\mathrm{S^{ts}},\mathrm{W}) = \mathrm{RL}(\mathrm{S^{ts}},\mathrm{W}) - \mathrm{IL}(\mathrm{S^{ts}},\mathrm{W}) \qquad (17)$$

An absolute value of RCL it called the clarity index (Equation (19)):

$$\mathrm{CL}(\mathrm{S^{ts}},\mathrm{W}) = |\mathrm{RL}(\mathrm{S^{ts}},\mathrm{W}) - \mathrm{IL}(\mathrm{S^{ts}},\mathrm{W})| \qquad (18)$$

The range of the clarity index is [0,1]. As it is mentioned, for a target trial the ideal value of RL is 0 and IL is 1, thus the ideal value of CL is 1. For a non-target trial, the ideal value of CL is also 1. Thus a higher value for the clarity index means that the decision is more dependable. Therefore we use it to select a sparse number of classifiers and use it in the weight learning process.

3.2.2 Weight Learning and Ensemble Selection

By using the clarity index in the classifier selection we solve three problems. The first problem is as a result of the fact that classifiers have different performance with respect to different test samples. This in fact, affects both classifier selection and weight determination, which is not exploited in previous works. The second problem is in choosing an efficient number of classifiers and the third is the correlation between the classifiers. There may be a different number of efficient classifiers for different test samples and they may or may not correlate in different situations. In sparse classifier fusion for speaker verification, these problems are not efficiently addressed either. By using the clarity index and a proper threshold value, we can choose an adaptive number of efficient classifiers. The proper threshold value is chose from the training scores so that the final EER for the training scores is minimized. In the case where the clarity index of all classifiers falls below the threshold, we use a predefined minimum number of classifiers.

In the ensemble selection process we do not have the weight vector to calculate the clarity index for each classifier. Thus, we change RL and IL formulation and replace $W^T S_i^{ntr}$ and $W^T S_j^{ptr}$ with s_i^{ntr} and s_j^{ptr} respectively. $s_i^{ntr}, i = 1, \dots, n_n$ are negative scalar scores and $s_i^{ptr}, i = 1, \dots, n_p$ are positive scalar scores related to the classifier.

We use two strategies for sample based classifier ensemble selection. In the first scenario, we choose a fixed threshold on the clarity index. Classifiers with a clarity index higher than the threshold are used in the ensemble. With this strategy, different classifiers are used for different test samples. In the case where all the indices are lower than the threshold, all 12 classifiers participated in the ensemble. In the second strategy, the threshold is not fixed and varies according to the values of the clarity index, related to each test sample. In this scenario, the test score of each classifier is first calibrated to log the likelihood ratio [19] then we use the threshold to select confident classifiers.

To take the sample specific behavior of classifiers and gain the generalization ability of equation (15) we propose to use equation (20):

$$\underset{\mathrm{W}}{\operatorname{argmin}}\, \mathrm{C_{wlr}}(\mathrm{W}) + \lambda\mathcal{F}(\mathrm{S^{ts}},\mathrm{S^{tr}}) \qquad (19)$$

Where $\mathcal{F}$ is a function of the test sample, current weights, and positive and negative training samples. If we directly substitute CL into the equation (20), the optimization becomes generally intractable, because due to the definition of RL and IL it is a discrete measure and cannot be differentiated. Thus we approximate the discrete relevant and irrelevant losses by differentiable sigmoid functions (Equations (21) and (22)):

$$\mathrm{RL}(\mathrm{S^{ts}},\mathrm{W}) = \frac{1}{\mathrm{n_n}}\sum_{\mathrm{i=1}}^{\mathrm{n_n}} \frac{1}{1 + \mathrm{e}^{-\alpha \mathrm{W^T}(\mathrm{S_i^{ntr}} - \mathrm{S^{ts}})}} \qquad (20)$$

$$\mathrm{IL}(\mathrm{S^{ts}},\mathrm{W}) = \frac{1}{\mathrm{n_p}}\sum_{\mathrm{j=1}}^{\mathrm{n_p}} \frac{1}{1 + \mathrm{e}^{-\beta \mathrm{W^T}(\mathrm{S^{ts}} - \mathrm{S_j^{ptr}})}} \qquad (21)$$

By choosing the correct value for α and β these two equations can be close to the original values of RL and IL. Setting a high value for α and β results in a closer approximation to the true values of RL and IL, but, results in several local optima for CL. On the other hand, a low value of these two parameters may result in a poor approximation of CL. Consequently these parameters have considerable effect on the performance of the classification. Even with this modification we still do not substitute CL into equation (20) because it is an absolute value of difference between RL and IL, and is not differentiable at zero. To get rid of this we use the ridge regression [16] of the raw clarity index (RCL).
Finally we define the optimization problem as:

$$\underset{\mathrm{W}}{\operatorname{argmin}}\, \mathrm{C_{wlr}}(\mathrm{W}) - \lambda\|\mathrm{RCL}\|_2^2 \qquad (22)$$

Although the optimization process is performed in the test phase, it is fairly fast and in less than half a second converges to the optimal points. This problem can be solved using standard packages [16]. For the case of faster optimization, we propose to optimize weights for all possible combinations of classifiers, and use proper weights after the ensemble selection using CL. Using this method, at first, we determine all combinations of classifiers. If we have n classifiers, the number of these combinations is $\sum_{i=1}^{n}\frac{n!}{(n-i)!i!}$. There are 4095 combinations when we have 12 base classifiers. Then, equation (24) is solved for each combination separately to obtain a table of weight vectors:

$$\underset{W}{\operatorname{argmin}}\, C_{wlr}(W) - \lambda\|w\|_2^2 \qquad (23)$$

In this equation, ridge regularization keeps weights small.

In the test process the score of each classifier is first calibrated. Then the clarity index is computed for each calibrated score, and confident classifiers are selected using the clarity index. Finally, confident scores are fused using related weights. Figure 2 depicts the block diagram of proposed ensemble classification system.

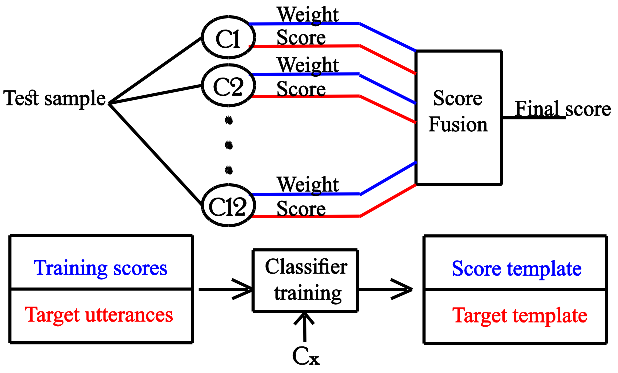


Fig. 2. Block diagram of the proposed ensemble classification system. Each of the classifiers C1-C12 contain score templates.

4. Experimental Results

4.1 Databases

We used NIST 2004 Speaker Recognition Evaluation (SRE), and switchboard II in our experiments. Since we use many classifiers and each classifier or feature has an ability to detect special characteristics of a speech signal, we preferred not to restrict training or test data to originating from a male or female, or specific language. NIST 2004 contains 6244 training files. The Universal background model is trained on these data. This dataset also contains 660 male and female speakers, and 4623 test utterances. These utterances are from five different languages: Arabic, English, Mandarin, Russian and Spanish. In the case in which an utterance has more than one minute duration we split the utterance to have more test data. Switchboard II (2348 conversation sides) is also used to train the PLDA dimensionality reduction process, λ, and nuisance attribute projection (NAP).

4.2 Experimental Setup

It is believed that diversity of base classifiers improves the performance of ensemble classification [20]. In addition, features used and the methods of classification should be efficient enough for the classification. Therefore, our experiments are conducted on three well-known different classifiers and four different feature vectors. We used MFCC, PLP, SWLP, and PLCC as different speech features.

Table 1. Twelve different base classifiers implemented on NIST04 dataset, using four different features and three methods plus fusion systems. Proposed spample based (Instance basced) methos spesified as IBSparse1&2

	Classifier	Feature	Devset			Evaset		
			EER(%)	MinDCF ×100	MinCLLR ×100	EER(%)	MinDCF ×100	MinCLLR ×100
1	Ivector-PLDA	MFCC	6.68	4.68	21.78	8.01	5.59	24.59
2	Ivector-PLDA	LPCC	7.25	5.74	23.39	5.79	4.85	19.44
3	Ivector-PLDA	PLP	6.59	5.06	21.59	7.76	5.27	23.92
4	Ivector-PLDA	SWLP	9.12	8.59	30.73	10.65	8.37	34.48
5	GMM-SVM-BHAT	MFCC	7.23	7.23	25.30	8.12	7.76	24.76
6	GMM-SVM-BHAT	LPCC	8.35	6.01	25.27	7.59	6.16	23.35
7	GMM-SVM-BHAT	PLP	8.15	6.67	25.08	7.71	6.30	22.01
8	GMM-SVM-BHAT	SWLP	10.54	8.19	30.53	11.05	8.48	29.54
9	GMM-SVM-KL	MFCC	7.44	5.53	23.78	9.12	6.71	26.94
10	GMM-SVM-KL	LPCC	6.66	4.69	23.98	8.41	5.67	25.40
11	GMM-SVM-KL	PLP	7.45	6.42	25.02	8.37	6.91	25.27
12	GMM-SVM-KL	SWLP	7.88	5.56	27.45	9.47	6.57	29.07
13	Sparse_fusion	-	**3.05**	**2.75**	**10.36**	3.37	3.02	11.63
14	IBSparse1	-	-	-	-	**2.56**	2.26	**8.01**
15	IBSparse2	-	-	-	-	2.89	**2.25**	9.72

At first energy based voice activity detection is performed on each utterance. Then, feature extraction is performed using a 25ms hamming window with 50% overlap (12.5ms). Voicebox MATLAB toolbox [1] is employed to extract MFCC features. To obtain PLP

[1] Available online at: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.zip

features, we used RASTAMAT MATLAB toolbox which is available online[1]. SWLP features, which we have used, are briefly described in [21], and there are MATLAB codes available online[2], to extract these features. To extract LPCC features, we used msf_lpcc.m MATLAB implementation[3]. We choose a feature vector dimension of 14 for all four features.

We used a 2048 Gaussian mixture model to create a universal background model on the training part of NIST 2004 speaker recognition evaluation corpus. MSR toolbox is used to extract all GMM models, and i-vectors [22]. GMM_UBM_KL and ivector_PLDA are implemented using MSR toolbox and voicebox. The i-vector dimensionality was 400, which is reduced to 200 using PLDA. An important matter in score fusion is that, scores from different classifiers may vary significantly, as a result of using different feature vectors and classification methods. Using results of [2] we used z-cal(clipped) as pre-calibration method.

To evaluate base classifiers and fusion method we considered EER, minDCF and minCLLR using BOSARIS MATLAB toolkit [23].

4.3 Results

In this section we consider three methods to fuse individual scores. In the first method we use a weighted logistic regression cost, C_{wlr}, regulated by E-net ($\alpha = 0.1$) [2]. In the second method we replace the regularization term with our proposed sample specific term, and perform ensemble selection using the clarity index (IBSparse1). In this method optimization is performed on each test sample. In the last experiment we calculated weight vectors for all possible combinations of the ensemble set (IBSparse2). We empirically found that best results can be obtained when the size of the ensemble is limited between 4 and 8 classifiers. In this situation, most suitable classifiers are selected based on the test sample and the optimization of weights is not performed in the test stage.

Table 1 shows that different classifiers have instance based behaviors. For example ivector-PLDA which uses PLP features, has the best EER for the development set, while this is not suitable for the evaluation set. The next evidence is the whole performance of GMM-SVM-KL in comparison to which is good with respect to other classification methods and is worse for the evaluation set. Comparing the performance of GMM-SVM-KL using LPCC features and ivector-PLDA using MFCC features supports the same idea.

As an example of performance improvement, we observed, sparse classifier fusion [2] results in the score of 5.1484 for the verification of the utterance '*xalm.sph*' which belongs to NIST SRE 2004, and class 1 (the first model in the database). Because this is a target score, it is better to be higher. The clarity index for this trial is as follows:

[0.845,0.555,0.825,0.66,0.355,0.64,0.935,0.53,0.72,0.87,0.98,0.875]

The proposed method chooses the 6 most confident classifiers which are: 1st, 3rd, 7th, 10th, 11th and 12th classifiers (of Table 1). The Fusion of scores of these classifiers results in a fused score of 7.2706.

To use the clarity index in ensemble set selection, a threshold value should be used. Values higher than the threshold are considered as confident classifiers and lower values are considered as belonging to unconfident classifiers. We obtained the threshold value from the development set and used it in the evaluation set for classifier ensemble selection. A comparison of three speaker verification fusion systems is shown in Figure 3. This curve is obtained using MSR MATLAB toolbox. It is clearly observed that the proposed method 1 shows the best results in almost all parts of the plot, with the cost of optimization of weights in the test phase.

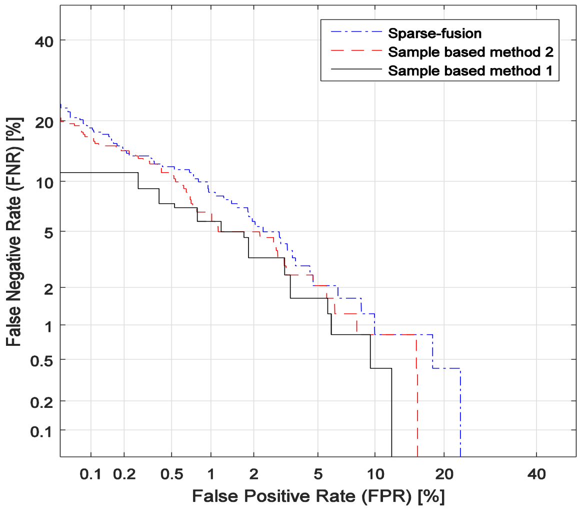


Fig. 3. DET plot of three speaker verification fusion systems (plotted using MSR MATLAB toolbox)

4.4 Correlation of Classifiers

Diversity of classifiers is an important issue in ensemble classification. A More diverse set of classifiers increases the chance of taking more aspects of the classification in to account. Correlation is the opposite point of diversity. One can use one of the two highly correlated classifiers without significant reduction in performance. In the case of our experiment we indirectly take the correlation into account. As it is mentioned in subsection 3.2.1, if the score of a classifier is less than all non-target scores, the classifier confidently tells that test sample does not belong to the claimed class and if the value is greater than all the target scores the classifier confidently tells that test sample does belong to the claimed class. In both the situations CL value is 1. Therefor higher values of CL belong to confident classifiers and lower values belong to unconfident ones. By choosing a proper threshold value of for the clarity we omit very unconfident classifiers. Therefore the remaining classifiers are assumed to be confident enough to

[1] Available online at: http://labrosa.ee.columbia.edu/matlab/rastamat
[2] Available online at: http://users.spa.aalto.fi/jpohjala/xlp/
[3] Available online at: https://github.com/jameslyons/matlab_speech_features/archive/master.zip

participate in the ensemble. A question that is raised here is what happens if some of the classifiers are highly correlated. For example, if seven classifiers are in the ensemble and four of them are highly correlated, it means they exploit the same speech characteristics and lower the effectiveness of other uncorrelated classifiers. If the weight of every classifier is fixed for all test samples, this effect reduces the performance of the ensemble, due to the fact that classifiers have different correlations for different samples. This issue remains while weight learning is performed in the development phase, including our sample specific fusion method 2. But when weights are learned in the test phase, even if the mentioned four classifiers are in the ensemble set, and exploit the exact same characteristics of the utterance, the weight learning algorithm gives them the most efficient weights.

5. Conclusions

We introduced a sample specific classifier fusion for speaker verification, which selects an adaptive number of best classifiers and determines sample specific fusion weights for each selected classifier. The method implements a group of well-known base classifiers for speaker verification, and ranks them using information obtained from labeled samples and individual unlabeled samples. The weight learning process uses logistic regression and the optimization problem is constrained with a sample specific term.

Extensive experiments on unconditioned, large variant NIST 2004 demonstrated the effectiveness of the proposed method. It would be interesting to perform experiments about the weight of constraint (λ) and the timing of the optimization formula.

References

[1] X. Zhou, A. S. d’Avila Garcez, H. Ali, S. N. Tran, and K. Iqbal, “Unimodal late fusion for NIST i-vector challenge on speaker detection,” Electron. Lett., vol. 50, no. 15, pp. 1098–1100, 2014.

[2] H. L. Ville Hautamä ki, Tomi Kinnunen, Filip Sedlák, Kong Ail Lee, Bin Ma, “Sparse Classifier Fusion for Speaker Verification,” IEEE Trans. Audio, Speech Lang. Process., vol. 21, no. 8, pp. 1622–1631, 2013.

[3] Z. Lei, Y. Yang, and Z. Wu, “Ensemble of support vector machine for text-independent speaker recognition,” IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 6, no. 5, pp. 163–167, 2006.

[4] A. Kumar and B. Raj, “Unsupervised Fusion Weight Learning in Multiple Classifier Systems,” arXiv:1502.01823, Feb. 2015.

[5] K. Lai, D. Liu, S. Chang, and M. Chen, “Learning Sample Specific Weights for Late Fusion,” Image Process. IEEE Trans., vol. 24, no. 9, pp. 2772–2783, 2015.

[6] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, “Temporally weighted linear prediction features for tackling additive noise in speaker verification,” IEEE Signal Process. Lett., vol. 17, no. 6, pp. 599–602, 2010.

[7] F.I. Cabeceran, “Fusing prosodic and acoustic information for robust speaker recognition,” Ph.D dissertation, TALP Research Center, Speech Processing Group, Universitat Politècnica de Catalunya Barcelona, July 2008.

[8] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O’Shaughnessy, “Multitaper MFCC and PLP features for speaker verification using i-vectors,” Speech Commun., vol. 55, no. 2, pp. 237–251, Feb. 2013.

[9] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, “SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation,” 2006 IEEE Int. Conf. Acoust. Speech Signal Process. Proc., vol. 1, no. 2, pp. 1–3, 2006.

[10] C. You, K.-A. Lee, and H. Li, “GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition,” IEEE Trans. Audio, Speech Lang. Process., vol. 18, no. 6, pp. 1300–1312, 2010.

[11] P. Mat, M. Kara, and P. Kenny, “full-covariance ubm and heavy-tailed plda in i-vector speaker verification,” Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, Prague, Czech Republic, 22 May - 27 May 2011 pp. 4516–4519, 2011.

[12] A. Solomonoff, W. M. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia PA, USA, 2005, pp. 629–632.

[13] P. Emerson, Designing an All-Inclusive Democracy. Springer, 2007.

[14] W. M. Campbell, D. E. Sturim, W. Shen, D. A. Reynolds, and J. Navrátil, “The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Honolulu, Hawaii, USA, 15-20 April 2007.

[15] S. Chernbumroong, S. Cang, and H. Yu, “Genetic Algorithm-based Classifiers fusion for multi-sensor activity recognition of elderly people,” IEEE J. Biomed. Heal. informatics, vol. 19, no. 1, pp. 282 – 289, 2014.

[16] C. M. Bishop, Pattern recogintion and machine learning (Information Science and Statistics). Springer, 2007.

[17] S. Pigeon, P. Druyts, and P. Verlinde, “Applying logistic regression to the fusion of the NIST’99 1_speaker submissions,” Digit. Signal Process., vol. 10, no. 1–3, pp. 237–248, 2000.

[18] M. Schmidt, G. Fung, and R. Rosales, “Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches,” Lect. Notes Comput. Sci., vol. 4701, pp. 286–297, 2007.

[19] N. Brümmer, “Measuring, refining and calibrating speaker and language information extracted from speech,” Ph.D dissertation, Faculty of Engineering, University of Stellenbosch, 2010.

[20] S. Wang and X. Yao, “Relationships between diversity of classification ensembles and single-class performance measures,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 206–219, 2013.

[21] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, “Stabilised weighted linear prediction,” Speech Commun., vol. 51, no. 5, pp. 401–411, 2009.

[22] S. O. Sadjadi, M. Slaney, and L. Heck, “MSR Identity Toolbox v1. 0: A matlab toolbox for speaker-recognition research,” Speech Lang. Process. Tech. Comm. Newsl., 2013.

[23] N. Brümmer and E. de Villiers, “Bosaris toolkit [software package],” 2011. [Online]. Available: https://sites.google.com/site/bosaristoolkit/.

Mohammad Hasheminejad received the B.Sc. degree in Bio-electrical engineering from University of Isfehan, Isfehan, Iran, in 2003. He received the M.Sc. degree in communication engineering from Maleke ashtar university of technology, Tehran, Iran, in 2008. He is currently Ph.D student in Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran. His area research interests include Image Processing and retrieval, Pattern recognition, Digital Signal Processing and Sparse representation. His email address is: mhashemi@birjand.ac.ir.

Hassan Farsi received the B.Sc. and M.Sc. degrees from Sharif University of Technology, Tehran, Iran, in 1992 and 1995, respectively. Since 2000, he started his Ph.D in the Centre of Communications Systems Research (CCSR), University of Surrey, Guildford, UK, and received the Ph.D degree in 2004. He is interested in speech, image and video processing on wireless communications. Now, he works as associate professor in communication engineering in department of Electrical and Computer Eng., university of Birjand, Birjand, IRAN. His Email is: hfarsi@birjand.ac.ir.

The Surfer Model with a Hybrid Approach to Ranking the Web Pages

Javad Paksima
Department of Engineering, Payame Noor Yazd University, Yazd, Iran
Paksima@pnu.ac.ir
Homa Khajeh*
Department of Engineering, Science and Art University, Yazd, Iran
khajeh121@yahoo.com

Received: 31/Oct/2015 Revised: 19/Jun/2016 Accepted: 26/Jul/2016

Abstract

Users who seek results pertaining to their queries are at the first place. To meet users' needs, thousands of webpages must be ranked. This requires an efficient algorithm to place the relevant webpages at first ranks. Regarding information retrieval, it is highly important to design a ranking algorithm to provide the results pertaining to user's query due to the great deal of information on the World Wide Web. In this paper, a ranking method is proposed with a hybrid approach, which considers the content and connections of pages. The proposed model is a smart surfer that passes or hops from the current page to one of the externally linked pages with respect to their content. A probability, which is obtained using the learning automata along with content and links to pages, is used to select a webpage to hop. For a transition to another page, the content of pages linked to it are used. As the surfer moves about the pages, the PageRank score of a page is recursively calculated. Two standard datasets named TD2003 and TD2004 were used to evaluate and investigate the proposed method. They are the subsets of dataset LETOR3. The results indicated the superior performance of the proposed approach over other methods introduced in this area.

Keywords: Ranking; Web Pages; Surfer Model; Learning Automata; Information Retrieval.

1. Introduction

The content of the World Wide Web is increasingly growing. According to the studies reported in [1],[2], there exist more than 1 billion websites on the Web. In this regard, search engines are considered efficient tools used for recovering and extracting important information from this large set of data. Mostly, about 91% of users use search engines to find their desired information [3]; moreover, users stated that 73% of the information provided by search engines was valid [3]. Without appropriate ranking, search engines are not able to meet users' needs. Users' tendency to find the result of query at the high ranks of the ranking list indicates the importance of ranking algorithms efficiency. Ranking means placing relevant pages at the first rank so that users can find the answers to their queries in the shortest possible time.

Ranking methods fall into two general classes, namely, content-based and connection-based. Content-based methods use the content of webpages. Instances of such methods include models which are Boolean, probability (like BM25 method [4]) and vector (like TF-IDF method[1] [5]). These methods suffer from rank spamming [6]. Rank spamming means that the owners of some webpages usually add extra and irrelevant words, which are mostly invisible and blended in the background color, to their pages to be more selected by search engines. In connection-based methods, webpages are evaluated using other pages. Links indicate the quality of destination page from the perspective of source pages. Instances of connection-based methods are PageRank [7], and HostRank [8]. The main problem of such methods is called Rich-Get-Richer [9]. This problem is caused when search engines always place popular pages at the top of the list resulting from users' queries, and users usually visit the first ten results. Therefore, the popular pages become more popular, and the new relevant pages are less likely to be visited. Hybrid approaches have been proposed to solve this problem. They are based on connection and content. For instance, HITS [10] and TSPR [11] use content to improve ranking. The proposed method is also a hybrid approach.

Connection-based algorithms are divided into two main categories including query-independent and query-dependent methods. In query-independent methods such as PageRank and HostRank, ranking is done using the entire web graph offline. However, in query-based methods such as HITS, ranking is done only in a part of web graph which includes the query-related pages. In [7], out-link uniformly is chosen at random to determine the page to visit at the next time step on the graph, but in this paper, chosen page with respect to non-uniform distribution The proposed method is query-dependent, too. Thus, ranking is done among the query-related pages.

[1] TF-IDF as the vector-space model is utilized in page ranking and this ranking method is named TF-IDF.

* Corresponding Author

The intelligent surfer, proposed in this paper, would not select pages with respect to uniform distribution. However, it selects them with respect to their contents and connections. For this issue, the learning automata would be used to calculate the probability of selecting pages. A page was selected to hop by the learning automata along with the contents and connections of pages in each phase. Moreover, the surfer could also select a page for transition from the current page with respect to the content of the connected pages. While the surfer is moving about webpages, their ranking is selected recursively.

The rest of this paper is organized as follows: Part 2 reviews the research literature, which focuses on the studies which deal with only the connection and content of webpages. A definition of the learning automata, used in the proposed method, is presented and the proposed method is introduced in Part 3. Then the evaluation criteria are discussed in Part 4, and the results are investigated. Finally, the conclusion and suggestions for future studies are presented.

2. Literature Review

A review of the research works related to ranking webpages is presented here. Generally, ranking algorithms are trying to study the following problems: What criterion can be used to indicate whether the webpage is related to users' query or not? How are the results ranked so that they respond to user's query at any time? What algorithm is able to place more relevant results at first ranks? Ranking algorithms are classified with respect to the use of content and connection. Vector space and probabilistic models are among the most important content-based methods.

Salton introduced a method for ranking the documents corresponding queries. In this model, document and query are considered vectors; their length is equal to the number of words existing in the term. Cosine of the angle between the two vectors is considered the degree of their similarity [12]. Salton et al. [5] proposed a model named TF-IDF in 1988. This method used the frequency of document and query words to calculate the weight. The idea used in this method is that the most frequently repeated word describes the document better, and the words appearing in fewer documents have more information. The advantage of this vector model is its simplicity and flexibility; however, there was no official framework to find and display the degree of proposed relationship.

The probabilistic model is one of the content-based models whose objective is to find the probability of dependency of each document on each query. Unlike the vector model, it cannot find the similarity degree definitely. Robertson [4] proposed BM25, which is one of the best probabilistic methods. In BM25, weighting is considered to be based on an okapi. This highly accurate probability formula indicates the similarity between queries [4]. This method suffers from rank spamming. PageRank algorithm is a query-independent method, in which, ranking value of each page is equal to the weighted summation of its input pages' ranks. In other words, a page has a high ranking if many pages refer to it, or the referring pages have high rankings [7]. TSPR method is the developed version of PageRank, which considers N headlines in the entire Web. Using PageRank, the pages are ranked with each headline. Content methods retrieve the pages containing query words, and PageRank score is calculated on different topics [11].

Ghodsnia and Yazdani [13] proposed a penalty and reward-ranking algorithm named BPRR, which added a new dimension to PageRank model through the direct feedback from user. The visiting priority of search results was considered a vote for this document, and the score values were attributed to documents [13]. This method reduces the impact of Rich-Get-Richer problem. In PPR (Personalized PageRank), data-mining technology is used to extract user's automatic interests [14]. The proposed algorithm moves gradually towards user's interests to personalize ranking. The common filter is used when a new query is made in order to improve ranking accuracy and validity.

WordRank method [15] is similar to PageRank approach. Their difference is that the user waltzes through pages in the same way as the random surfer does in PageRank [15]. However, the user does not select the external links to a page with equal probabilities in this method; rather, he operates as a directed surfer. The user selects a page similar to the current page.

The WeightedPageRank method [16] is similar to PageRank. Their difference is that both the internal and external links are considered [16]. TrustRank [17] is a method to cope with rank spamming. It is a semi-automatic method to distinguish between good and bad pages. Good pages are the trusted ones (in which rank spamming did not occur). The idea of TrustRank algorithm is that when a page with trust degree of one is distributed to other pages, the impact of trust degree is reduced as the distance is increased [17].

Pandey [18] proposed a method by which a balance would be established between new quality pages and the available ones [18]. Using this method, new pages have the chance to be placed at the top of the ranking list. In HostRank method, the pages are first placed in a hierarchical structure of host directory, named the superior group, to obtain the connections on the graph. Then, the degree of each node is distributed among the pages containing the node by the hierarchical structure [8]. This method solves the problems of excessive distribution of webpages and Rich-Get-Richer.

In [19],[20], a ranking method named RL_Rank was proposed with a query-independent approach. In this algorithm, the user is a random surfer who moves between webpages. Moving between webpages is done by clicking on one of the external links of the current page. A reward is considered for the selected page, and the value function of each page indicates its rank. The results showed the superiority of the proposed method over PageRank.

Zareh et al. [21] introduced an algorithm based on reinforcement learning named DistanceRank. In this method, the logarithm distance between pages was used as the received reprimand, and the objective is to minimize the total received reprimand. This algorithm is less sensitive to Rich-Get-Richer phenomenon [21]. In query-independent algorithms, all pages are competing, and irrelevant pages sometimes rank higher due to their popularity. HITS algorithm [10] is executed on a sub graph named root. It is then expanded to a bigger graph named the base graph. For each vector such as v, two variables of a(v) and h(v) named authority and hubness are defined. It means that a page of high authority is referred to by a number of pages with high hubness, and a page of high hubness refers to a number of high authorities [10]. SALSA [22] is a connection-based method. The idea of this method is to combine PageRank with HITS. It creates a repetitive graph between hubs and authority. At first, some nodes are haphazardly selected out of authority nodes. Then the values of variables are adjusted with repetition between previous and next steps [22].

Due to the problems existing in the two types of algorithms (connection-based and content-based), hybrid approaches were proposed to increase the accuracy of such algorithms. Shakeri and Zhai [23] introduced a hybrid ranking method, in which, a score named hyper-relevance was defined for each page. The rank of each page is calculated through the linear combination of three parameters such as the similarity between page and query, weighted functions for internal and external links [23]. The most important disadvantage of this model is that it should be online, a problem which reduces the response time. Zareh et al. [24] used the click data to combine some content and connection features of webpages. The simulated click data is used to allocate the weight to each feature; therefore, the best combination would be found [24].

In [25], a method from the combination of two methods; Enhanced-RatioRank and Page level keyword, is proposed. This method uses the concept of structure and keyword search at a page level. The problem which can be seen in method [25] is that each page must be clicked by the user once. This is possible for personal dictionaries or websites, but it will not be possible for web pages in a search engine. Method [26] is a developed model of PageRank, which forms some tables from the keywords of pages, and by considering the similarity between the titles and the user's query words, accords greater importance to a particular subject. Their aim is to personalize the ranking for the user, and they use the browser history data. Whereas, the method proposed in this paper can be used for the query of any user, and there is no need to receive information from the client side except the query phrase to rank pages by a search engines.

PageRank uses a model based on the random surfer agent to rank webpages, and selects one of the pages with an external link for transition at a uniform probability. Otherwise, it jumps to another page in a uniform way. An intelligent surfer is proposed in this paper. The selection of pages for hopping or transition is not based on the uniform distribution. The pages are selected with respect to their contents and connections. The learning automata were used to calculate the probability of selecting pages so that it could hop to other webpages. The contents features of referring pages were used for transition to one of the pages of the external link. The proposed method is query-dependent.

3. The Proposed Method

The idea of the proposed method is to create an intelligent surfer that moves between the retrieved webpages for query. This method considers the relationships between pages and their contents. The surfer has two approaches to transition from the current page to other pages. First, it goes to one of the linked pages. Then, it hops on one of the retrieved webpages. Given this reasoning, if the linked page is linked to another page, it can be stated that the page topic was appealing to the creator of the first page. Therefore, the link indicates the interest of another page in this page. The surfer hops on one of the pages to which it is linked. This page is selected with respect to the contents of the referring pages.

BM25 was particularly used as the best content feature in this paper. The reason the contents of referring pages were used is to consider the concept that relevant pages point to each other, and a relevant page can be relevant in terms of content. In the second case, the surfer considers the page rank, which is updated during movement in the graph with a source of content feature. The connection feature of webpage calculates the probability of selecting for hopping by using the learning automata. Exploring the webpages, the surfer calculates their ranks recursively. Put it another way, ranking webpages is converged to constant values. In the following statement, ranking score of web pages is calculated as section 3-1. Moreover, a definition of the learning automata, used in the proposed method, is presented in section 3-2. After that a page probability is calculated by Learning Automata and MB25 for proposed method as section 3-3. At last, section 3-4 presents proof of proposed method convergence.

3.1 Calculation Ranking of Web Page

In the proposed method, if the intelligent surfer is on page i, it hops on one of the pages related to external links or it hops on other pages with respect to their probabilities. The likelihood of selecting external links will not be the same. The surfer selects a link, which is more relevant in terms of content. The probability of a transition from the referring page to the referred page is equal to BM25 of the referred page divided by the summation of BM25's of all of the external links (Unlike reality, since it is possible that all of the external links are irrelevant in terms of the content feature of BM25, a very slight amount of ε was added to all BM25's to prevent the denominators from being zero.) Therefore, the score which page j receives due to a transition from page i to page j (with respect to its external link) is equal to $\frac{R_{iq}(k)\times(BM25_{jq}+\varepsilon)}{\sum_{z\in O(i)}(BM25_{zq}+\varepsilon)}$, which is calculated recursively.

If the external links are not selected, the surfer hops on another page that will be selected with respect to probability. The page rank is calculated through the following equation:

$$R_{jq}(k+1) = (1-d)\times P_{jq}(k+1) + d \times \sum_{i\in B(j)} \frac{R_{iq}(k)\times(BM25_{jq}+\varepsilon)}{\sum_{z\in O(i)}(BM25_{zq}+\varepsilon)} \quad (1)$$

In which $R_{jq}(k+1)$ and $P_{jq}(k+1)$ indicate the rank and probability of page j for the query q in the $(k+1)^{th}$ step, respectively. $O(i)$ is equal to a set of pages of external links pertaining to page i. $B(j)$ indicates a set of pages referring to page j. d represents the damping factor. Parameter d is used to guarantee the convergence of the proposed method and to delete the impact of sink pages (the ones with no external links). $BM25_{zq}$ indicate content feature of BM25 of page z for query q. ε is the constant value of 0.05. All of the pages compete with each other with the same topic, and increase accuracy. This method would decrease the Rich-Get-Richer problem. The learning automata would be used to calculate the page probability. Description of the learning automata is used as follows:

3.2 Background of the Learning Automata (LA)

The idea of the learning automata was stated by Testlin in the early 1960. A learning automaton [27],[28] operates in a random environment. It is a comparative decision-making unit, which selects the optimal action out of a set of finitely authorized operations through repeated interactions with learning. It improves the efficiency, and the action is selected haphazardly. The selected action is the input of the random environment at each moment. The environment responds to the action with a reinforcement signal. The probability vector is updated through the reinforcement feedback from the environment. The aim of learning automata is to minimize the average penalty received from the environment. A learning automaton is useful for an environment with insufficient information [29]. A learning automaton is also well appropriate for an environment, which is complicated, dynamic and random with uncertainty. The reason for that is the use of learning automata in a wide range of issues such as optimization problems [30], computer network [31], grid computing [32], signal processing [33], information retrieval [34], and Web engineering [35].

The environment can be defined with {α,β,c} in which α≡{α1,α2,..,αr} indicates a finite set of inputs, while β≡{β1, β2,.., βm} refers to a set of variables which can be selected, and c≡{c1,c2,…, cr} represents a set of penalty probabilities in which ci depends on the given action αi. If the penalty probability is constant, the previously mentioned random environment turns to a constant random environment, and if it changes with time, the environment is named an inconstant one. Depending on the nature of reinforcing signal β, the environments can be categorized as p-model, S-model and Q-model. In P-model environments, the reinforcing signal can only be two binary values (0 or 1). In Q-model, a finite number of values ranging in [0, 1] can be selected for the reinforcing signal. In s-model, the reinforcing signal is in [a, b].

The learning automata can be divided into two main classes [27]: the learning automata with a constant structure and that with a changing structure. The first one is indicated with <β, α, L> in which β is the set of inputs while α is the set of actions, and L is the learning automata. The learning algorithm is used to change the probability vector. It allows αi(k)ϵα and p(k) actions to be selected by the learning automata. The probability vector is defined for this set of actions at the moment of k. The parameters of reward and penalty are indicated with a and b. The number of actions that the learning automata can select is indicated with r. At each constant of k, the probability vector of p(k) is updated through the linear algorithm resulting from Eq. (2). If the selected activity of αi(k) is the reward given by the random environment, the updated penalty results from Eq. (3):

$$p_j(k+1) = \begin{cases} p_j(k) + a\left[1-p_j(k)\right] j = i \\ (1-a)p_j(k) \forall j \neq i \end{cases} \quad (2)$$

$$p_j(k+1) = \begin{cases} (1-b)p_j(k) j = i \\ \left(\frac{b}{r-1}\right) + (1-b)p_j(k) \forall j \neq i \end{cases} \quad (3)$$

If a=b, Equations (2) and (3) are named the reward of linear penalty (L_{R}-$_{p}$). If a>>b, the above equations are named the reward of linear penalty ($L_{R\text{-}\epsilon P}$). However, if b=0, they are named the reward of linear inactivity ($L_{R\text{-}I}$). In the last case, if the action is fined by the environment, the probability vectors of the remaining action do not change [36].

The World Wide Web includes thousands of web pages, there is not sufficient information about which pages are relevant to the user query, and such information can only be obtained by surveying the environment. In this paper, learning automata is used as a tool that easily explores an environment about which there is no knowledge, and that acquires the required knowledge by interacting with the environment. This interaction is done by surveying the web graph, and the knowledge is obtained by updating the probability vector.

At each step, it selects one of the actions based on the action probability vector. This way of selection is based on the knowledge obtained from the environment, and it is better than selecting one of them randomly. The required knowledge is obtained from the web graph (environment).

3.3 Calculation Page Probability using LA

While calculating the page probability, it was assumed that the retrieved pages formed the state space of the learning automata, and the number of learning automata actions was equal to the number of the retrieved pages, which are more than the threshold, except for the current page. Threshold is equal average double of HostRank and average double of pages rank in pervious step. In each cycle, only the pages with higher ranks from the previous step and with respect to HostRank are selected. The condition which should be met to select the page is as follows:

$$\mathrm{HostRank_i} > \mathrm{HostRank_{avg}} * \rho \ \text{ and } \mathrm{R_i} > \mathrm{R_{avg}} * \rho \tag{4}$$

In which R_i and $HostRank_i$ indicate the rank of page *i* which was calculated with the proposed method and HostRank [8], respectively. $\mathrm{HostRank_{avg}}$ and $\mathrm{R_{avg}}$ indicate the average values of HostRank for the retrieved pages for user's query and the average rank of the retrieved pages for user's query in the previous step, respectively. ρ is the constant value of 2. With coefficient 2, a top quarter of the ranked list is considered to be relevant. In this state, better results have been provided.

The probability is calculated according to Eq. (3). The probability of the selected page increases (it is rewarded), and those of other pages decrease. In Eq. (1), $P_j(k+1)$ represents the probability of page *j* in cycle *k+1* while *a* indicates the reward according to the following relation.

$$\mathrm{a} = \mathrm{e}^{-\frac{\beta(\mathrm{T}-\mathrm{t})}{\mathrm{t}}} \tag{5}$$

In which, β is the constant value of 4.4 known as the step size. T indicates the entire number of execution cycles for convergence. It is assumed that the set of possible actions in each step is equal to the number of retrieved pages pertaining to user's query. In each step, if the page rank in the previous step is greater than twice of the average page ranks pertaining to the query, and the HostRank of the page is greater than twice of average HostRank of pages pertaining to the query, a reward will be received. The page ranks are calculated recursively. Finally, the pages are arranged in a descending order with respect to their ranks. The pseudo code and Module pertaining to the proposed algorithm are indicated in Figs. 1 and 2, respectively. Table 1 shows the parameters used in Fig. 1. Moreover, the convergence of the proposed method is empirically proved in the next part.

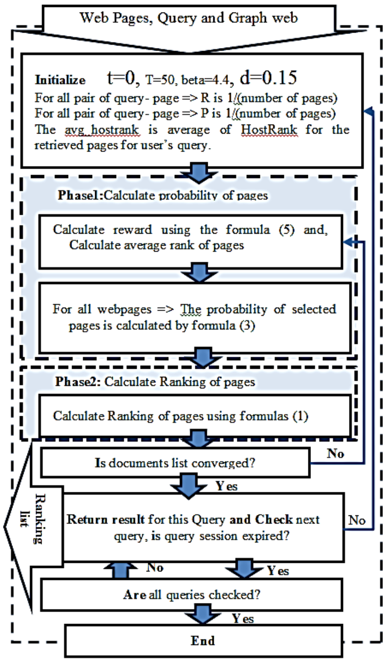
Table 1. Parameters used in the pseudo-code of Proposed Method

Parameter name	Represent
N	The total number of pages retrieved for query
t	The number repeats or time
T	The total number of execution cycles for convergence.
Beta	It is the constant value of 4.4 known as the step size.
D	It is the damping factor.
$\mathrm{R_{iq}}$	rank of page i for the query q
$\mathrm{P_{iq}}$	probability of selected page i for the query q
avg_hostrank	The average values of HostRank for the retrieved pages for user's query.
avg_R	The average rank of the retrieved pages for user's query.
$\mathrm{ParentCount_i}$	The number of members in the set of pages pointed to page i.
Sum_val_links	The sum of BM25 of out-links.
$\mathrm{LinkCount_p}$	out-degree of the page p

Algorithm

```
Input
graph_link
n: number of pages retrieved for query q^th
list of document-query pair for query q^th
Output
R: Ranking list
Initialize
t=0, T=50, beta=4.4, d=0.15;
For i=0 to n
 R_iq=1/n;  //R used for Rank of pages
 P_iq=1/n;// P used for probability of pages
end
avg_hostrank = averge$_{i\in\{0,..,n\}}$ $HostRank_i$
Begin
For t=0 to T
Phase1:
//Calculate probability of pages
 a=exp(-beta*(T-t)/T);
avg_R= averge$_{i\in\{0,..,n\}}$ $R_{iq}$
 For i=0 to n
  if($R_{iq} > 2 * avg\_R \ and \ \ HostRank_i > 2 * avg\_hostrank$) then
           P_iq=P_iq + a*(1-P_iq)
  else
           P_iq=(1-a)*P_iq
   end
  end
Phase2:
//Calculate Ranking of pages
  For i=0 to n
    Value=0
  For p=0 to parentCount_i
      Sum_val_links=0
    For j=0 to linkCount_p
   Sum_val_links = Sum_val_links+BM25_jq+$\varepsilon$
      end
        Value=Value+ R_pq/ Sum_val_links
      end
        R_iq= (1-d)*P_iq + d*(BM25_iq+ $\varepsilon$ )*Value
   end
end
```

Fig. 1. The Proposed Method Pseudo Code



Web Pages, Query and Graph web
Initialize t=0, T=50, beta=4.4, d=0.15
For all pair of query- page => R is 1/(number of pages)
For all pair of query- page => P is 1/(number of pages)
The avg_hostrank is average of HostRank for the retrieved pages for user's query.
Phase1:Calculate probability of pages
Calculate reward using the formula (5) and, Calculate average rank of pages
For all webpages => The probability of selected pages is calculated by formula (3)
Phase2: Calculate Ranking of pages
Calculate Ranking of pages using formulas (1)
Is documents list converged?
No
Yes
Ranking list
Return result for this Query and Check next query, is query session expired?
No
No
Yes
Are all queries checked?
Yes
End

Fig. 2. Module of the proposed algorithm

3.4 Proof of Convergence

The convergence of the proposed method is empirically proven in this part.

A similarity test was conducted to prove the convergence of the proposed method empirically. The aim of ranking was to sort the pages out with respect to their scores [37]. Therefore, the results of different iterations are compared with each other to display the convergence. For this purpose, the similarity resulting from the 20th repetition was compared with a sorted list including 5th, 8th, 10th, 11th, 15th, 16th, 17th, and 19th repetitions. The similarity of two lists is calculated according to the following equation:

$$\text{Similarity} = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

In which A and B contain the N first pages on the sorted list resulting from two different repetitions. $|A \cup B|$ indicates the total number of pages which appeared on the N first pages of the two lists (list union), and $|A \cap B|$ indicates the number of pages which appeared on the N first pages of both lists (list intersection). Fig. 3 shows the similarity among different repetitions in comparison to the 20th repetition for 2 to 47200 pages. This test was conducted on the dataset TD2003. If the similarity of two lists resulting from two repetitions gets close to 1, it means that the list of pages proposed in two repetitions was almost the same and the order of pages were constant after these repetitions. In other words, the ranking order has converged.

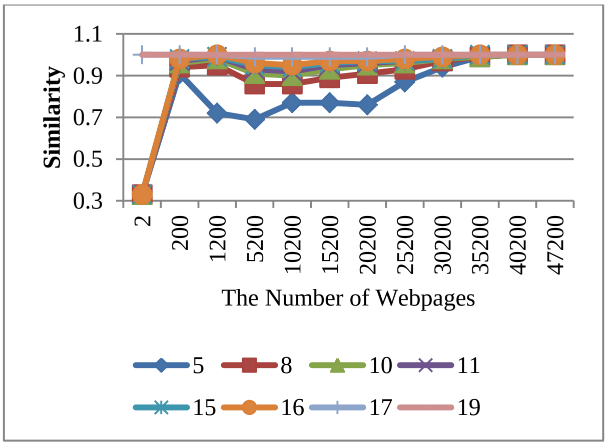


Fig. 3. The Convergence of the Proposed Method during the Successive Repetitions of Execution

4. Assessment of Empirical Results

In this part, the empirical results of evaluating the efficiency of the proposed algorithm on TD2003 and TD2004 datasets, taken from the standard set of ranking algorithms test LETOR [38], is presented.

Evaluation was made between the proposed method and the previous ones, using standard criteria such as MAP, P@n, and NDCG@n [39].

4.1 Assessment Criteria

In this evaluation, precision at the position of n (P@n), Mean Average Precision (MAP) and Normal Discount Cumulative Gain (NDCG) were used to evaluate the efficiency because they are comprehensively used in information recovery. They are defined as follows:

❖ **P@n**

Precision at n calculates the relevance of n webpages at the top of the list of ranking results with respect to a query.

For instance, if the first ten retrieved documents by the query are as {relevant, irrelevant, irrelevant, relevant, relevant, relevant, irrelevant, irrelevant, relevant, irrelevant}, P@1 to P@10 are as {1, 1.2, 1.3, 2.4, 3.5, 4.6, 4.7, 4.8, 5.9, 6.10}, respectively.

Precision at n and the relevance of n documents at the top of ranking list are calculated with respect to user's query [38].

$$P@n = \frac{\#\text{ of relevance docs in top n results}}{n} \tag{7}$$

❖ **MAP**

For an average query, precision is defined as the average P@n values for all relevant documents. The average precision (AP) [38] is calculated through the following equation. It is equal to the average value of P@n for all relevant documents.

$$AP = \frac{\sum_{i=1}^{N}(P@i * rel(i))}{\#\text{ total relevance docs for this query}} \tag{8}$$

In which N indicates the number of retrieved documents, and rel(n) shows the binary function. If the nth document is relevant, it is 1; otherwise, it is zero. Finally, MAP is equal to the average precision (AP) for all queries [38].

❖ **NDCG@n**

It is the evaluation criteria of the cumulative gain which have been normalized. It represents the judgment on the multi-level relevance. The value of NDCG of ranking list at position n is calculated through the following relation for a query:

$$NDCG(n) = Z_n \sum_{j=1}^{n} \frac{2^{r_j}}{\log(j+1)} \tag{10}$$

In which r_j is the degree of relevance for document *j* on the ranking list. Z_n is the normalization constant, which is determined in a way that the highest value of NDCG would be one. For LETOR3, there are two degrees for relevance, {0 and 1}, from user's perspective. They indicate irrelevance and relevance, respectively [38].

❖ **NWN**

Another point states the most accuracy of the ranking methods on different datasets [38]. They propose a metric called Winning Number to evaluate the performance of ranking methods over the datasets included in the LETOR 3.0 collection. Winning Number is defined as the number of other algorithms which is better than they are. The Winning Number [40] is calculated according to Eq. (11).

$$\mathrm{WN_i(M)} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \sum_{\mathrm{k=1}}^{\mathrm{m}} \mathrm{I_{\{M_i(j)>M_k(j)\}}} \tag{11}$$

In which, n and m are the number of datasets and algorithms in the comparison, respectively. *j* indicates the index of a dataset, *i* and *k* are indices of an algorithm, *M* is an assessment criterion (such as MAP or NDCG), $M_i(j)$ represents the performance of the i^{th} algorithm on the j^{th} dataset, and $I_{\{M_i(j)>M_k(j)\}}$ indicates an indicator function such that

$$\mathrm{I_{\{M_i(j)>M_k(j)\}}} = \begin{cases} 1 & \text{if } \mathrm{M_i(j) > M_k(j)} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

The Winning Number evaluation metric depends on the denseness of the evaluation results. This means that there were evaluation results for all rank algorithms on all datasets under comparison [41]. Therefore, the normalized Winning Number metric is proposed to enable comparison of a sparse set of evaluation results. This Normalized Winning Number takes each dataset into account, and an algorithm is evaluated on and divides this by the ideal Winning Number [41]. The indicator function I is defined in order to only take into account datasets on which it has been evaluated. If $\mathrm{M_i(j)}$ and $\mathrm{M_k(j)}$ are both defined and $\mathrm{M_i(j) > M_k(j)}$ is true, it is 1; otherwise, it is zero.

The Normalized Winning Number is calculated according to the following equation:

$$\mathrm{NWN_i(M)} = \frac{\mathrm{WN_i(M)}}{\mathrm{IWN_i(M)}} \tag{13}$$

In which, $\mathrm{IWN_i}$ is the Ideal Winning Number and, i is index of $\mathrm{i^{th}}$ algorithm. IWN theoretically is equal to the highest Winning Number.

4.2 Benchmark Datasets

Similar settings and conditions were required to evaluate the efficiency of the proposed method and to compare it with other approaches. Standard benchmark datasets of TD2003 and TD2004 were used in this paper. They were published at LETOR website [39] for the same purpose.

In addition, tests were carried out on a computer with an Intel i7 core 2.10 GHz CPU and 6 GB memory.

4.3 Empirical Results

Benchmark datasets of TD2003 and TD2004 were used to evaluate the proposed method. The results were stated with respect to the evaluation criteria of P@n, MAP, and NDCG@n. The proposed algorithm was compared with algorithms such as BM25, HostRank, PageRank, and HITS. In the figures, HITS_a and HITS_h meant HITS based on authority and hub, respectively. The proposed method has been called alg_automata in the figures. According to Figs. 4 to 5, the proposed algorithm showed a better performance on two benchmark datasets of TD2003 and TD2004 with respect to the evaluation criteria of P@n, MAP, and NDCG@n. It is noteworthy that Table (1) indicates the improvement percentage of the proposed algorithm in comparison to HostRank, BM25, HITS_h, HITS_a, and PageRank algorithms with respect to P@n, MAP, and NDCG@n. The highest improvement percentage was observed in three criteria on TD2003 compared to PageRank and on TD2004 compared to HITS_h.

The evaluation of empirical results of the proposed method can be seen in Figs. 4 to 5. They indicate the efficiency of the proposed method on two datasets of TD2004 and TD2003. They also showed the superiority of the proposed method over other algorithms. The proposed method worked better than PageRank because PageRank is in favor of the old pages, and new pages do not have many links, even if they are really good.

Table 2. The Improvement Percentage of the Proposed Method Compared to Other Algorithms

	Dataset	TD2003			TD2004		
	Evaluation Criteria	P@n	MAP	NDCG@n	P@n	MAP	NDCG@n
Algorithm	HITS_h	%124.82	%124.18	%165.84	%267.63	%161.71	%267.81
	HITS_a	%69.07	%55.13	%69.41	%47.21	%42.96	43.53%
	HostRank	%48.71	%61.52	%56.89	%14.30	%18.92	%17.42
	BM25_title	%54.49	%31.48	%51.	%66.46	%34.50	%60.48
	PageRank	%163.82	%148.38	172.37%	%80.57	62.33%	%78.57

Considering the contents of pages, the proposed method does not have this problem. It reduced Rich-Get-Richer problem. In PageRank, popular pages tend towards general popularity; however, the popularity of website is not guaranteed by taking enough information. Considering the content in the proposed method, this problem is solved; therefore, a better ranking is provided. The proposed method was better than HITS because of the following two reasons. First, HITS suffers from topic drift. It means that if irrelevant pages exist in the root set with strong connections, these irrelevant pages are reflected on the pages in the basic set. Moreover, the web graph is made up of the webpages of the basic set which will not have more relevant nodes, and the results of algorithm will not be able to find pages with high hubs and authorities for the query. The second reason is that HITS considers the same value for links although it may not provide user's query with the relevant topic. The proposed method considers a value for each link with respect to the content of pages, and this will result in its superiority. Compared to BM25, the proposed method also pays attention to the link between pages. This would also result in its superiority. It has been considerably improved over HostRank because of paying attention to the contents of pages.

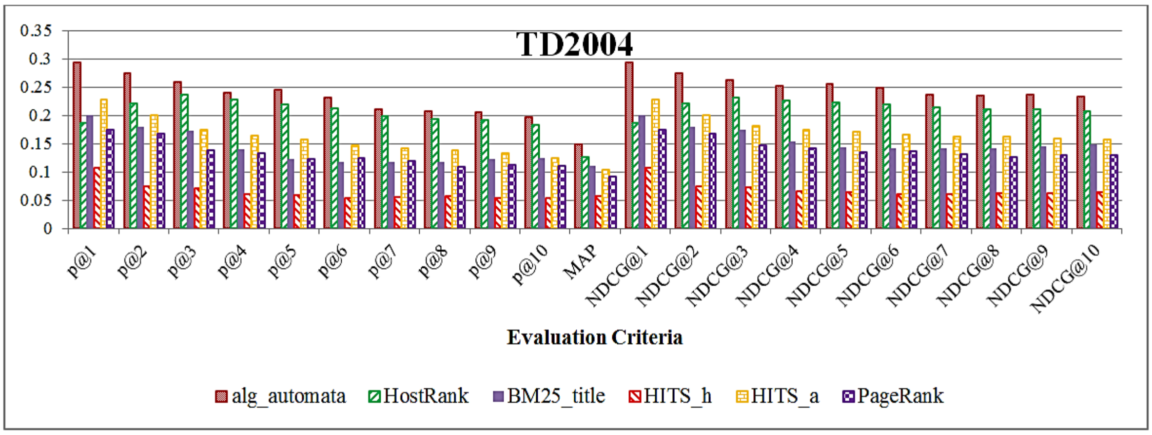


Fig. 4. Comparing the Proposed Method with Other Algorithms on TD2004

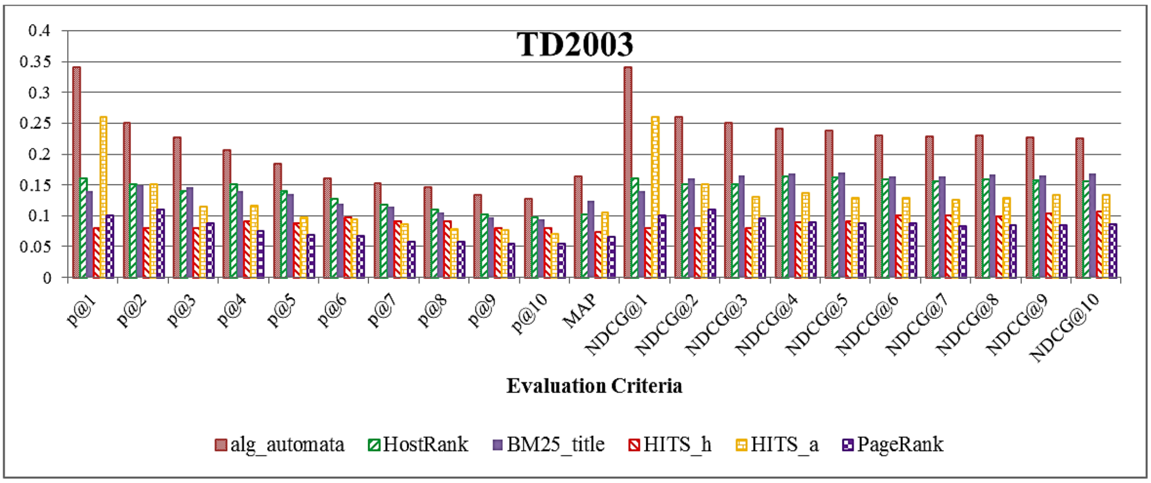


Fig. 5. Comparing the Proposed Method with Other Algorithms on TD2003

P@n, NDCG@n and MAP are the used evaluation metrics in the used datasets combined (for $n \in \{1,\ldots,5\}$). Fig. 6 shows NWN as function of IWN for the considered methods in this paper. The proposed method scores very high NWN scores on two datasets on MAP, P@n and NDCG@n (for $n \in \{1,\ldots,5\}$). HITS_a performed the NWN, around 0.8, on two datasets and also, performed well in both certainty and accuracy. BM25 is one of the best performers in the MAP comparison with a reasonable number of benchmark evaluations. There is a slight certainty on the accuracy of HITS_h and PageRank as both methods are evaluated on the two datasets included in the comparison for the NWN metric.

Fig. 7 shows the WN of methods based on results of MAP, P@n, and NDCG. The proposed method scores an IWN of 10 on two datasets, which is achieved by obtaining the highest score on the LETOR 3.0 TD2003 and TD2004 in this paper. HITS_a has a low WN value. The WN score of HITS_h is about zero. The WN and NWN scores for HITS_h and PageRank are lower than those for the other ranking methods, the certainty of their ranking performance is considered to be lower.

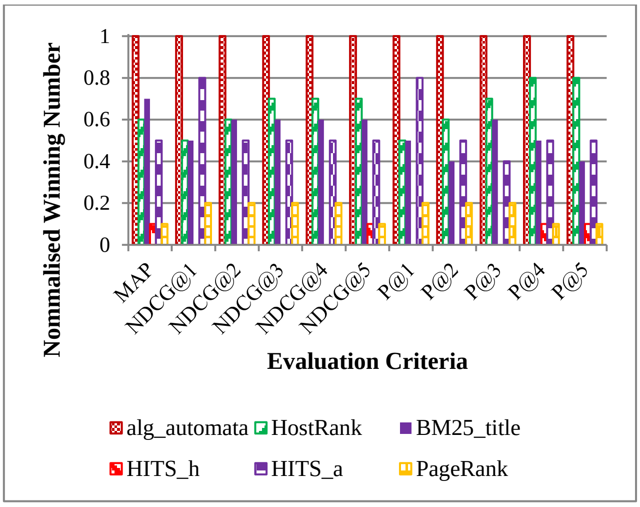


Fig. 6. Comparing the Proposed Method with Other Algorithms by Respect to Evaluation Criteria of NWN on TD2003 and TD2004

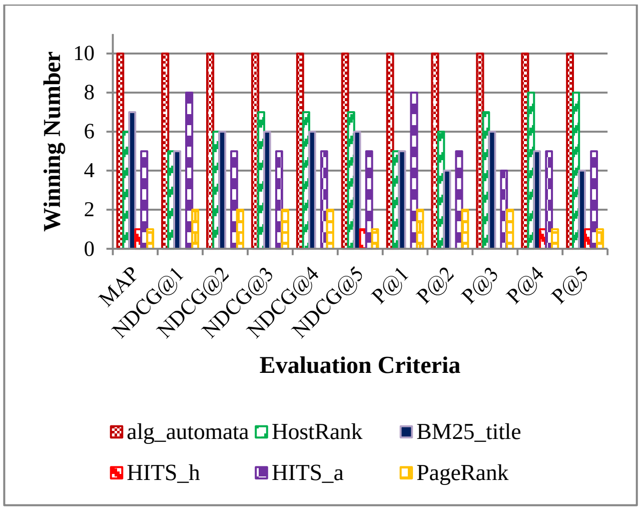


Fig 7. Comparing the Proposed Method with Other Algorithms by Respect to Evaluation Criteria of WN on TD2003 and TD2004

5. Conclusion and Future Suggestions

The proposed ranking method introduced an intelligent surfer that selects the pages with respect to their probability values. To calculate the probabilities, it was assumed that the retrieved pages were in the form of learning automata, and each page indicated a status. The number of actions of each learning automata was equal to the number of pages retrieved, except for the current page. Therefore, each page will have the chance of selection. Put it another way, the random surfer can hop to each page. Pages were selected for transition with respect to their scores, which were. This score was allocated to the page based on its content and connect. LETOR3 benchmark datasets, two standard datasets of TD2003 and TD2004 in particular, were used for evaluation. The empirical results indicated that the proposed method had better efficiency in comparison to content-based, connection-based, and hybrid methods such as BM25, HITS, PageRank and HostRank on TD2003 and TD2004 with respect to the evaluation criteria of P@n, MAP, and NDCG. The proposed method provided the users with the results of their queries due to being query-dependent. This method is based on content and connection; therefore, the proposed method could decrease the impact of problems such as rank spamming and Rich-Get-Richer. The proposed method has no other method superior to them in both IWN and NWN.

Using the methods of calculating probability of uniform distribution as the probability of hopping to webpages was postponed to the future along with the use of reinforcement learning methods.

References

[1] R. Albert, H. Jeong, and A.-L. Barabási. “Internet: Diameter of the world-wide web,” Nature, vol. 401, no. 6749, pp. 130–131, Sep. 1999.

[2] S. Lawrence and C. L. Giles. “Accessibility of information on the web,” Nature, vol. 400, no. 6740, p. 107, Jul. 1999.

[3] K. Purcell, J. Brenner, and L. Rainie. (2012, Oct.). “Search Engine Use 2012,” PEW Research Center, [Online]. p. 42, 2012 Available: http://www.pewinternet.org/2012/03/09/search-engine-use-2012/. [Sep. 1,2014].

[4] S. E. Robertson, and S. Walker. "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in Proc. 17th annual international ACM SIGIR conf. Research and development information retrieval, Springer-Verlag New York, Inc, 1994, pp. 232-241.

[5] G. Salton and C. Buckley. “Term-weighting approaches in automatic text retrieval.” Information processing and management, vol. 24, no. 5, pp. 513–523, Dec. 1988.

[6] M. R. Henzinger, R. Motwani, and C. Silverstein. “Challenges in web search engines.” In ACM SIGIR Forum, vol. 36, no. 2, pp. 11–22, Sep. 2002.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. “The PageRank Citation Ranking: Bringing Order to the Web.” World Wide Web Internet Web Information System, vol. 54, no. 1999–66, pp. 1–17, Jan. 1998.

[8] G.-R. Xue, Q. Yang, H.-J. Zeng, Y. Yu, and Z. Chen. “Exploiting the hierarchical structure for link analysis,” in Proc. 28th annual international ACM SIGIR Conf. Research and development in information retrieval, 2005, pp. 186–193.

[9] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks.” Science, vol. 286, pp. 509–512, Oct. 1999.

[10] J. M. Kleinberg. “Authoritative Sources in a Hyperlinked Environment.” Journal of the ACM (JACM), vol. 46, pp. 668–677, Sep. 1999.

[11] F. Qiu and J. Cho. “Automatic identification of user interest for personalized search,” in Proc. 15th int. conf. World Wide Web, 2006, pp. 727–736.

[12] G. Salton. The SMART retrieval system—experiments in automatic document processing. Amsterdam: IOS Press, 1971.

[13] P. Ghodsnia, A. M. Z. Bidoki, and N. Yazdani. “A punishment/reward based approach to ranking,” in Proc. 2nd int. conf. Scalable information systems, 2007, p. 58.

[14] W.-C. Peng and Y.-C. Lin. “Ranking web search results from personalized perspective,” in E-Commerce Technology, 2006. 8th IEEE Int. Conf. Enterprise Computing, E-Commerce, and E-Services, 3rd IEEE Int. Conf., 2006, p. 12.

[15] A. Kritikopoulos, M. Sideri, and I. Varlamis, “WordRank: A Method for Ranking Web Pages Based on Content Similarity,” presented at Databases, 2007. BNCOD’07. 24th British Nat. Conf., 2007, pp. 92–100.

[16] W. Xing and A. Ghorbani, “Weighted pagerank algorithm,” Commun. Networks and Services Research, 2004. in Proc.. 2nd Annual Conf., 2004, pp. 305–314.

[17] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. “Combating web spam with trustrank,” in Proc. 30th Int. Conf. Very large data bases-Vol. 30, 2004, pp. 576–587.

[18] S. Pandey, S. Roy, C. Olston, J. Cho, and S. Chakrabarti. "Shuffling a stacked deck: the case for partially randomized ranking of search engine results," in Proc. 31st Int. Conf. Very large data bases, 2005, pp. 781–792.

[19] E. Khodadadian, M. Ghasemzadeh, V. Derhami, and S. A. Mirsoleimani, "A novel ranking algorithm based on Reinforcement Learning." presented at 16th CSI Int. Symposium on Artificial Intelligence and Signal Processing (AISP 2012), 2012, pp. 546–551.

[20] V. Derhami, E. Khodadadian, M. Ghasemzadeh, and A. M. Zareh Bidoki. "Applying reinforcement learning for web pages ranking algorithms." Applied Soft Computing, vol. 13, no. 4, pp. 1686–1692, Apr. 2013.

[21] A. M. Zareh Bidoki and N. Yazdani. "DistanceRank: An intelligent ranking algorithm for web pages." Information Processing and Management, vol. 44, no. 2, pp. 877–892, Mar. 2008.

[22] R. Lempel, and S. Moran. "SALSA: the stochastic approach for link-structure analysis." ACM Transactions on Information Systems (TOIS), vol. 19, no. 2, pp. 131-160, Apr. 2001.

[23] A. Shakery and C. Zhai. "Relevance Propagation for Topic Distillation UIUC TREC 2003 Web Track Experiments." in TREC 2003, 2003, pp. 673–677.

[24] A. M. Z. Bidoki and J. A. Thom. "Combination of documents features based on simulated click-through data." in Advances in Information Retrieval, Springer Berlin Heidelberg, 2009, pp. 538–545.

[25] L. Rodrigues and S. Jaswal. "An Efficient Page Ranking Approach Based on Hybrid Model." presented at Adv. in Computing and Communation Engeering (ICACCE), 2015 2nd Int. Conf., 2015, pp. 693–696.

[26] T. H. Haveliwala. "Topic-Sensitive PageRank." in Proc. 11th Int. Conf. World Wide Web, Www2002.Org, 2002, pp. 517–526.

[27] K. S. Narendra and M. A. L. Thathachar. Learning automata: an introduction. Courier Corporation, 2012.

[28] S. Lakshmivarahan and M. A. L. Thathachar. "Bounds on the convergence probabilities of learning automata." IEEE Transactions on Systems, Man, and Cybernetics, A Systems Humans, vol. 6, no. 11, pp. 756–763, Nov. 1976.

[29] E. Billard and S. Lakshmivarahan. "Learning in multilevel games with incomplete information. I." Systems Man, and Cybernetics Part B Cybernetics IEEE Trans., vol. 29, no. 3, pp. 329–339, Jun. 1999.

[30] J. A. Torkestani and M. R. Meybodi. "A New Vertex Coloring Algorithm Based on Variable Aaction-Set Learning Automata." Computing and Informatics, vol. 29, no. 3, pp. 447–466, Jan. 2012.

[31] N. Kumar, N. Chilamkurti, and J. J. P. C. Rodrigues. "Learning automata-based opportunistic data aggregation and forwarding scheme for alert generation in vehicular ad hoc networks." Computer Communications, vol. 39, pp. 22–32, Feb. 2014.

[32] M. Hasanzadeh and M. R. Meybodi. "Grid resource discovery based on distributed learning automata." Computing, vol. 96, no. 9, pp. 909–922, Sep. 2014.

[33] S. Bhattacharyya, A. Sengupta, T. Chakraborti, A. Konar, and D. N. Tibarewala. "Automatic feature selection of motor imagery EEG signals using differential evolution and learning automata." Medical & biological engineering & computing, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[34] J. Akbari Torkestani. "An adaptive learning to rank algorithm: Learning automata approach," Decision Support Systems, vol. 54, no. 1, pp. 574–583, Dec. 2012.

[35] J. A. Torkestani. "An adaptive focused web crawling algorithm based on learning automata." Applied Intelligence, vol. 37, no. 4, pp. 586–601, Dec. 2012.

[36] H. Beigy and M. R. Meybodi. "A mathematical framework for cellular learning automata." Advances in Complex Systems, vol. 7, no. 03n04, pp. 295–319, Sep. 2004.

[37] T. H. Haveliwala. (1999,Oct.). "Efficient computation of PageRank," Technical Report, Database Group, Computer Science Department, Stanford University, [On-line]. 1999. Available: http://ilpubs.stanford.edu:8090/386/ [Dec., 1, 2014].

[38] T. Qin, T. Y. Liu, J. Xu, and H. Li. "LETOR: A benchmark collection for research on learning to rank for information retrieval," Information Retrieval, vol. 13, no. 4, pp. 346–374, Jan. 2010.

[39] J. Xu, T.-Y. Liu and H. Li. "The Evaluation Tool in LETOR," Microsoft Research Asia [Online]. Available: http://research.microsoft.com/en-us/um/beijing/projects/letor/LETOR3.0/EvaluationTool.zip [Dec. 1,2014]

[40] T.-Y. Liu. Learning to Rank for Information retrieval, Springer Science & Business Media, 2011.

[41] D. H. T. N. Sander Bockting. "A cross-benchmark comparison of 87 learning to rank methods." Information processing & management, vol. 51, no. 6, pp. 757–772, Nov. 2015.

Javad Paksima received his B.Sc. and M.Sc. degree in computer engineering from Sharif University, Tehran, IRAN, in 1996 and 1998. He is currently an Instructor in the Department of Computer engineering, Payame Noor University, Tehran, IRAN. He is currently a student in computer engineering in Yazd University and also is a member of Parsijoo team (Persian Search Engine). His research interests are in the areas of Computer Networks, and Information Retrieval.

Homa khajeh received the B.Sc. degree in Software Engineering from Islamic Azad University, Najafabad Branch (IAUN), in Isfahan, Iran, in 2009 and her M.Sc. degree of Software Engineering from Science and Art University in Yazd, Iran, in 2014. Her research interests are mainly in the field of Information Retrieval, Search Engine, Machine Learning, and Big Data.