An Evolutionary Clustering Approach with FCM in Layering Scheme for Computer Security in Network

Seyed Mahmood Hashemi and Jingsha He

***Abstract*—Security management is a challengeable concept of computer system in distributed environment. Classical approaches do not adapt to circumstances, so developers can not design a computer system base on common approaches that be controlled with administrator. In this paper, a layering approach is presented for security management. Proposed approach clusters the users, so administrator can control relation for security. Proposed approach enriches with multi-objective evolutionary optimization algorithm. Multi-objective evolutionary optimization algorithms have dynamic process, so they can adapt to conditions of distributed environment.**

***Index Terms*—Layering security management, clustering, FCM, multi-objective evolutionary optimization, multi-objective simulated annealing.**

I. Introduction

Security is the important issue for users of computer system, so developers need an approach with their circumstances to design the computer systems. Security has three concepts: 1-confidentially 2-integrity 3-availability. Confidentially is the concealment information or resources. Integrity refers to the trustworthiness of data. Availability refers to the ability to use the information or resource desire. Classical methods are not responsible for requirements of new applications. Main reason for unreliability of older methods is their nature. Since older methods use absolute rules, they can not adapt new situations with dynamic conditions. The best way is using the approach with dynamic process. Another problem is occurring when application is used in distributed environment. Nodes in distributed environment become disabling and this is usual event. If nodes have security relations (access control) security relations will destroy. In this paper, we proposed an approach that enables an application to distribute the security relations between multiple nodes and define security relations with dynamic process. In presented approach, we cluster users with Fuzzy C-Mean (FCM) method. Request for access data send from each cluster to administrators, then administrators decide about request (accept/reject). Two parameters of FCM need to tune, so we use Multi-Objective Simulated Annealing (MOSA). Since Evolutionary Optimization (EO) algorithms have dynamic process, MOSA is able to conform to environment conditions. Users are identified base on their experiment.

Users are belonged to clusters, so it is easy for administrators to apply the security rules on them. Each cluster has a centroid. Since we use soft clustering, each user can belongs to multiple clusters. Data access request send from cluster to another layer which are data sources.

Proposed approach has two layers: one layer is for users and the other layer belongs to data sources. Administrator provides a relation between the layer of users and layer of data sources. In the distributed environment (network), each node is a data source. Administrator must judges about data access request from users. If administrator decides the request is valid, users can access to data source. Administrator decides independently for each request, so security relation does not destroy with disabling nodes. Therefore, proposed approach provides a scheme with two separate layer and administrator connects them to each other. Data sources are in one layer and users are in the other layer.

The rest of this paper is organized as follow: Section II is about related works; Section III is assigned to optimization and in Section IV we discuss about FCM; in Section V we present our approach; Section VI is conclusion.

II. Related Works

The increasing of web technologies with complexity in process is focused on [1]. Gerardo Canforda et al, represents a scenario which confidential data are more exposed unlawful disclosure, thus they propose a three level method for confidentiality. The highest level is represented by the privacy regulation (PR). The intermediate level is the set of private objectives (PO), which are semi-structured statements describing how data can be accessed by users. The formulation of POs depends on the entities, the particular relationships among them, and the specific domain dictionary. The lowest layer of the model is represented by the private rules set which implement a given PO. A rule assumes the form of a query that the users can or can not send to database. The three layer representing is also represented in [2]. An IoT (Internet of Things) system contains tree layers: a physical perception layer that perceives physical environment and human social life, a network layer that transforms and processes perceived environment data and an application layer that offers context-aware intelligent services in a pervasive manner. Layered approach can be used in different medium such as the radio. In [3], layered approach is used in radio network. The cognitive radio is based on the software defined radio with adjustable operational parameters. The software allows the radio to tune to different frequencies, power levels and modulation schemes to establish or maintain a communication link. The cognitive radio network also is further adaptable to change situation with its ability to operate

Manuscript received August 4, 2015; revised April 23, 2016.

The authors are with the Beijing University of Technology, School of Software Engineering, Beijing, China (e-mail: Hashemi2138@yahoo.com, jhe@bjut.edu.cn).

successfully in collaborate or uncooperative networks. Paper analyzes attacks and mitigation techniques for both scenarios. The threats are classified according to the protocol layer upon which the attack is performed: Physical layer, Data link layer, Network layer, Application layer and cross-layer. Cross-layer attacks are those in which the attack is launched utilizing one layer while the attack targets another layer. Another aspect, which is crucial same as confidentiality, is Trust Management. [4] Proposes a scheme for dynamic trust management in P2P networks. P2P networks have the potential of converting any host into a data server and to use it as a part of a large system for disseminating information without the limitation of using a single (host) interface. A peer user usually is interested to storing the downloaded file and most likely executes it. This process leaves a front door for viruses to the local host. Several interesting studies about proliferation have been presented. So they discuss the performance of the current P2P trust management strategy with consideration of internal file infection and show that file infection has the potential to underscore proliferation countermeasures. To bound virus proliferation, they propose the Double-layer Dynamic Trust (DDT) management scheme, which uses a two-layer trusting strategy aimed to alleviate the impact of the internal infection. There are number of researchers that use Artificial Intelligence tools for trust system. In [5], Trust and Reputation System (TRS) are proposed to identify trustful cooperators. Authors propose a novel and flexible Trust Computation Model (TCM) based on Artificial Neural Network (ANN) to quantify the trust relationships between agents. We propose a broker-assisting information collection strategy based on clustering method in order to improve the performance of the system. Trust of data can be examined with various approaches. For example, [6] use a graph for trust. Onion routing networks hide user's identities behind a circuit of selected onion routers. However, they run a high risk of being compromised in the presence of the adversaries who employ malicious onion routers to perform correlation-like attacks. Exiting trust-based onion routing computes trust only according to user's own knowledge. In this paper, a novel trust graph based onion routing that mitigates key limitations in the use of trust for protecting anonymity. SGor is designed based on two key insights: 1- if people can assign trust to others according to their own knowledge independently, the trust from a group of honest people is more likely to be correct than the trust from a single honest person. 2- Although users have no immediate knowledge for their unfamiliar routers, these routers are not necessarily controlled by adversaries. Data quality approaches may be used in different types of network. In [7], data quality, which is based on cross-layered, is used in Wireless Sensor Network. In many applications of wireless sensor network (WSN) contexts the location of sensor node is important information that can be used to identify the location of an event of interest. This paper, tackles both secure localization and privacy issues in order to define an integrated solution that consider a sound privacy management policy coupled with a secure localization protocol. The presented approach is based on the assessment of data quality, which are evaluated to which extent the information to be processed by application in reliability and trustworthiness. This is done by introducing a way to evaluate the overall data quality when several cheap protection techniques are combined together. Although none of the used techniques guarantee reliability and trustworthiness by itself, we exploit consistency across them to evaluate data reliability. As a result, we introduce a protocol, name cross-layer protocol (CLP) that defines fundamental steps for assessing data quality [8], [9].

A common method, which is used in various aspects of security, is Clustering. A large number of clustering algorithms exist, but it is difficult to find a single clustering algorithm to get well detection effect. Fanfei Weng *et al*., introduced a new clustering algorithm, the Evidence Accumulation (EA) for intrusion detection based on the concept of clustering ensemble. In this approach, K-mean algorithm runs N times (as number as data) to find appropriate cluster. In [10], is used to intrusion detection. The paper proposed one kind of k-means algorithm based on the k-medics cyclic method and the improved triangle trilateral relations theorem, which improves the k-means algorithm from reduce makes the improvement to the initial cluster center dependence and the algorithm time expenses. Eduardo Raul Hruschka *et al.*, presented a survey for evolutionary algorithms in clustering [9].

Researches, which are mentioned above suffers from a number of problems and they are focus on some notes. Firstly, they can not combine the rules for different aspects of security successfully. Secondly, adaptation between access demands and rules is problem. Papers provide some notes that must be used for contribution. At the first is layering. The second subject is clustering. This technique (clustering) causes perfect partitioning. In present paper, both of them are noted.

III. Multi-objective Optimization

In many fields, there is a need to set variables which cost function to be optimized. There are some approaches for optimization, but all of them can be categorized into two main groups: 1-deterministic 2-evolutionary. Deterministic Optimization Algorithms achieve optimize values for variables. Although deterministic algorithms have mathematical proof, they suffer long process time. Deterministic algorithms have final result after some iteration and before final iteration there is not any solution. Unfortunately, processes with deterministic time can not work in network environment. However Evolutionary Optimization (EO) Algorithms do not have mathematical proof to achieve best solutions for optimization problems, their time process can be controlled. EO algorithms have some loops and a number of loops can be controlled base on condition of system. Of course, EOA can not achieve best results in any iteration, but against deterministic approach they have a solution.

The major goal of EO algorithms is optimum value, but this goal is changed when there is a Multi-Objective Optimization (MOO) problem. The aim of MOO is tuning the decision variables to satisfy all objective functions F_i to optimum value. This class of problem is modeled by [10]

$$\text{Optimize} \qquad [F_1(X),...,F_K(X)]$$

$$\text{S.T.: } g_i(X) \le 0, h_j(X) = 0; i = 1, \cdots, m; j = 1, \cdots, p \quad (1)$$

where *K* is the number of objective functions, *X* is the decision vector, *m* is the number of inequality constraints and *p* is the number of equality constraints.

This goal causes differences between these algorithms and their ancestor single-objective optimization, which is based on concept of *best*, while the multi-objective optimization uses the concept of *dominance*. Dominance is defined in [10]:

$$\vec{U} = (u_1, ..., u_k) \prec \vec{V} = (v_1, ..., v_k)$$
$$\text{if } \forall i \in \{1, ..., k\} \Rightarrow u_i \le v_i, \exists j \in \{1, ..., k\} \Rightarrow u_j < v_j \quad (2)$$

In words, a vector $\vec{U}$ of variables dominates another vector of variables $\vec{V}$ if and only if $\vec{U}$ can reach to optimal value for some criteria without causing a simultaneous non-optimal value for at least one criterion. If two vectors cannot dominate each other, they are called as *non-dominated* vectors.

Multi-objective Simulated Annealing (MOSA)

Basic concept in Simulated Annealing is evolution of the solution by simulating the decreasing temperature (*tmp*) in a material, where higher the temperature meaning that higher the modification of the solution at a generation. If temperature of a hot material decreases very fast its internal structure may diverse and materials become hard and fragile. Decreasing temperature slowly yields higher homogeneity and less fragile materials. Evolution of the solution is carried at specific temperature profiles. At the first iterations a diverse set of initial solutions for the problem is produced at higher temperatures. And, these solutions are evolved while the temperature decreases to get their local optimums. In multi-objective situations, there are non-dominated solutions which must be kept in the archive, as a candidate of optimal solution.

Along the runs of MOSA algorithm, there are two solutions: current-so and new-so. They can have one of three states compared to each other: i- current-so dominates new-so, ii- current-so and new-so are non-dominated each other and iii- new-so dominates current-so.

If new-so is dominated by current-so, there may be solutions in archive which dominates new-so. New-so is accepted to the archive by the probability

$$p = \frac{1}{1 + \exp(\Delta \,.\, tmp)} \quad (3)$$

where Δ is differentiating between new-so and other solutions which dominates new-so

$$\Delta = \frac{\sum_i^k \Delta_i + \Delta}{k+1} \quad (4)$$

Solutions can escape from local-optima and reach to the neighborhood of the global-optima by this probable acceptance.

If new-so is dominated by some solutions in the archive, (4) is modified to:

$$\Delta = \frac{\sum_{i=1}^k \Delta_i}{k+1} \quad (5)$$

When new-so is non-dominated with all members in archive, then new-so is set as current-so and it is added to the archive.

If new-so dominates some solutions in the archive, then new-so is set as current-so and it is added to the archive and solutions in the archive which are dominated by new-so are removed.

If new-so is dominated by some solutions in the archive, then (3) is changed to:

$$p = \frac{1}{1 + \exp(-\Delta)} \quad (6)$$

where Δ is the minimum of the difference between new-so and dominating solutions in the archive. New-so is set as current-so with the probability (6). If new-so is non-dominated by all solutions in the archive it is set as current-so and added to the archive. If new-so dominates some solutions in the archive, it is set as current-so; it is added to the archive; and all dominated solutions are removed from the archive [11]-[13] (see Fig. 1).

1. Set *current-so*;
2. Produce *new-so*;
3. Compare current-so and new-so:
   - 3.1. IF *current-so* dominates *new-so* THEN *new-so* is accepted to archive with (3) and compare with other solutions in archive
     - 3.1.1.IF *new-so* dominate archive solutions THEN *current-so* is replaced with *new-so*
     - 3.1.2.IF archive solutions dominate *new-so* THEN *new-so* accepted in archive with (6)
     - 3.1.3.IF *new-so* is non-dominated with all archive solutions THEN *new-so* set as *current-so* and *new-so* add to archive
     - 3.1.4.IF *new-so* dominated some archive solutions THEN *new-so* set as *current-so* and dominated solutions remove from archive
   - 3.2. IF *new-so* dominates *current-so* OR *new-so* is non-dominate *current-so* THEN *new-so* set as *current-so*
4. Algorithm iterates until termination conditions;

Fig. 1. Pseudo code of MOSA.

IV. Fuzzy C-Mean (FCM)

The FCM algorithm scores each data vector $x_i = (x_{i,1}, \ldots, x_{i,k}) \in R^k$ in the data set $\{x_1, x_2, \ldots, x_N\}$ into *C* clusters according to a distance measure by solving the cost function [14]:

$$\min J_m(U, V) = (U, V)^m \, dist^2(x_i, v_a)$$
$$u_{a,i} \in [0, 1]; \ \forall a = 1, \ldots, C; \ \forall i \in I$$

$$\sum_{a=1}^{C} U_{a,i} = 1;\ 0 < \sum_{i=1}^{N} U_{a,i} < 1 \quad (7)$$

where $U=(u_{a,i}) \in R^{C.N}$ is the partition matrix, also called the fuzzy-membership matrix; $V = (v_{a,k}) \in R^{C.k}$ is the matrix of cluster centers, v_a is the center of a^{th} cluster; dist(x_i, v_a) is the distance between vectors x_i and v_a. The scalar $m>1$ is called *fuzzifier* or *fuzzification power*, and it determines the fuzziness of clustering. If m is closer to 1 then $U_{a,i}$ tends to crisp values {0, 1}, and, if m is large then $U_{a,i}$ tends to distribute gradually in interval [0, 1].

$$u_{i,a} = \begin{cases} \left(\sum_{j=1}^{C} \left(\frac{\| x_a - v_i \|}{\| x_a - v_j \|} \right)^{2/(m-1)} \right)^{-1} & ;if \ \| x_a - v_j \| > 0 \\ 1 & ;if \ \| x_a - v_i \| > 0 \\ 0 & ;if \ \exists j \neq i, \| x_a - v_j \| = 0 \end{cases} \quad (8)$$

where $a = 1, \ldots, N$ and $i = 1, \ldots, C$.

$$v_i = \frac{\sum_{a=1}^{N} u_{i,a}^m x_a}{\sum_{a=1}^{N} u_{i,a}^m} \quad (9)$$

Mostly Euclidian distance is preferred in clustering real data sets:

$$\text{dist}(X_a, V_i) = \left[\sum_{s=1}^{K} \left(x_{a,s} - v_{i,s} \right)^2 \right]^{1/2} \quad (10)$$

Getting the optimum solution for (7) is difficult [14]. A deterministic algorithm is proposed by some researchers [15], [16] to solve this optimization problem, which might fail to get the global optimum. An alternative solution for FCM algorithm is defined by [13]. Some researchers propose a method to specify appropriate numbers of clusters [15]-[17]. In [18], it is proposed that clusters shall provide the following two features: *minimum inside variance* (variance of vectors in that cluster) and *maximum outside variance* (variance between clusters). The other concept that must be satisfied by the clusters is maximization of the average of membership values. The average of membership values is calculated dividing sum of membership values of all data in a cluster by the number of data in that cluster [13]. Clusters which have low average of membership value are merged to the clusters to obtain higher average membership values. For this reason, authors proposed a formula for scoring the clusters:

$$S_i = \frac{\sum_{a=1}^{N} U_{i,a}}{N} \quad (11)$$

In [19], these aspects are declared in the other words and it is said *optimal partition* of data into subgroups were based on three requirements: (i) clear separation between resulting clusters; (ii) Minimal volume of clusters; (iii) Maximum number of data points concentrated in the vicinity of the cluster centroid. These aspects are defined on the concept of *partition density* which is defined by:

$$P_D = \frac{S}{\sum_{k=1}^{C} \left[\det(F_k) \right]^{1/2}} \quad (12)$$

where $S = \sum_{j=1}^{N} \sum_{k=1}^{C} U_{i,k}$,

$$F_k = \frac{\sum_{j=1}^{N} h(k \mid X_j) (X_j - V_k) (X_j - V_k)^T}{\sum_{j=1}^{N} h(k \mid X_j)} \quad (13)$$

and $h(k \mid X_j)$ is the probability of selecting the i^{th} cluster given the j^{th} feature vector. The average partition density is calculated from:

$$D_{PA} = \frac{1}{C} \sum_{i=1}^{C} \frac{S_i}{\sum_{k=1}^{C} \left[\det(F_k) \right]^{1/2}} \quad (14)$$

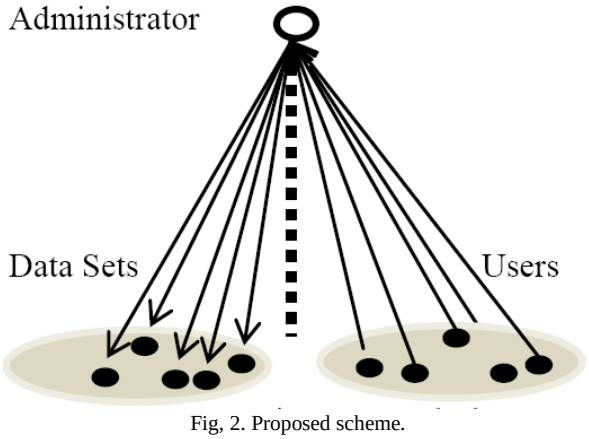
where $S_i = \sum_{j=1}^{N} U_{i,j}$.

In [17], these aspects redefined as similarity and dissimilarity between clusters.

V. Proposed Scheme

There are two problems in the common security approach. Firstly, they are not suitable in distributed environments. In distributed environments, the nodes may become disabling usually. If disable nodes had security relations, they destroy with disability of nodes. Since disabilities of nodes are very usual in distributed environments (networks), the method to keep security relations is necessary. Secondly, definition of security relation in dynamic process is necessary. Dynamic process allows to administrators for controlling the security.

The proposed scheme (see Fig. 2) consists two layers. One layer includes data sets and the other layer includes users. Users are clustered. The administrator connects two layers. Access demands (requests of access to data set) from cluster of users send to the administrator and if access demands follow the security relations, the administrator allows users to access to data sets.



Fig, 2. Proposed scheme.

Users are clustered with Fuzzy C-Mean (FCM). Each cluster is stored in an individual node.

This scheme has some features. Firstly, users are clustered, so the management of clusters is easier than the management of separate users. Secondly, since the scheme provides soft clustering for users, users may be belonged to multiple clusters. It causes, destroying especial clusters does not deeply affect on performance of system (see Fig. 3).

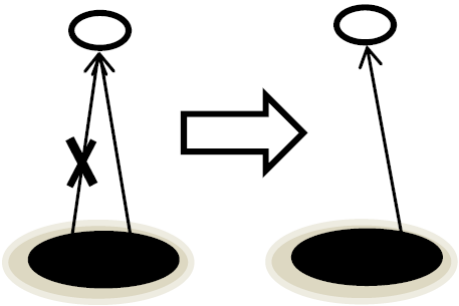


Fig. 3. A connection disability.

The main issue in the proposed approach is clustering. We use Fuzzy C-Mean (FCM) to cluster users. Actually, the administrator keeps the security relations with each user. FCM clusters users with the soft method, so users may be belonged to multiple clusters. There two parameters in FCM which need to define. We use Multi-Objective Simulated Annealing (MOSA) for the definition of FCM parameters.

Two parameters of FCM need to define: the number of clusters (C) and the power of fuzzification (m). Definition of FCM parameters is according to (6) and (8). Those formulas declare three conditions that are explained in the previous section. Optimization of C, m is doing with MOSA (see Fig. 4):

- Do
  - Set C, m as *current-so*
  - Do FCM with (8)
  - Calculate objectives with (12), (14)
  - Do MOSA algorithm
- Until termination condition

Fig. 4. Pseudo code of proposed algorithm.

After 30 epochs, follow values in Table I are produced for C and m:

TABLE I: RESULTS

C	m	Partition Density	Average Partition Density
5	3.9205	14.0723	1.4598

Since MOSA is stochastic, final results depend on MOSA parameters completely.

VI. CONCLUSION

In this paper, we propose an approach for the security management of computer systems. Users are clustered with Fuzzy C-Mean (FCM). Our method for clustering is soft clustering. It means, can belong to multiple clusters. The administrator manages the security with managing the cluster requests. Actually, managing the request of clusters is easier than managing the all of user requests. There are two parameters of FCM which need to tune. Multi-Objective Simulate Annealing (MOSA) tunes FCM parameters. Since MOSA is based on our approach and MOSA is a stochastic algorithm, different executions may have different final results.

REFERENCES

[1] G. Canfora, E. Costante, I. Pennino, and C. Aaron Visaggio, "A tree-layered model to implement data privacy policies," *Computer Standards & Interfaces*, vol. 30, pp. 398-409, 2008.

[2] Y. Zheng, Z. Peng, and A. V. Vasilakos, "A survey of trust management for internet of things," *Journal of Network and Computer Application*, vol. 42, pp. 120-134, 2014.

[3] D. Hlavacek and J. M. Chang. (2014). A layered approach to cognitive radio network security: A survey. *Computer Networks.* [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2014.10.001

[4] C. Lin and R. Rojas-Cessa, "Mitigation of malware proliferation in P2P networks using double-layer dynamic trust (DDT) management scheme," National Science Foundation under Grant Award 0435250.

[5] Z. Bo, X. Feng, J. Jun, and L. Jian, "A broker-assisting trust and reputation system based on artificial neural network," in *Proc. the 2009 IEEE International Conference on Systems, Man and Cybernetics*, 2009.

[6] Z. Peng, L. Xiapu, C. Ang, and R. K. C. Chang, "SGor: Trust graph based onion routing," *Computer Networks*, vol. 57, pp. 3522-3544, 2013.

[7] A. Coen-Porisini and S. Sicari, "Improving data quality using a cross layer protocol in wireless sensor networks," *Computer Networks*, vol. 56, pp. 3655-3665, 2012.

[8] W. Fangfei, J. Qingshan, S. Liang, and W. Nannan, "An intrusion detection system based on the clustering ensemble," *IEEE*, 2007.

[9] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, A. C. P. Leon, and F. de Carvalho, "A survey of evolutionary algorithms for clustering," *IEEE Transaction on Systems, Man and Cybernetics—Part C: Application and Reviews*, vol. 39, no. 2, 2009.

[10] L. Tian and W. Jianwen, "Research on network intrusion detection system based on improved K-means clustering algorithm," in *Proc. the 2009 IEEE 2009 International Forum on Computer-Science Technology and Applications*.

[11] W. L.-X. Wang, "A course in fuzzy systems and control," Prentice-Hall International Inc., pp. 118-127.

[12] C. A. Coello, D. A. Van Veldhuizen, and G. B. Lamont, *Evolutionary Algorithms for Solving Multi-objective Problems*, 2nd ed. Springer, 2007, pp. 30-45.

[13] S. Haojun, W. Shengrui, and J. Qingshan, "FCM-based model selection algorithms for determining the number of clusters," *Pattern Recognition*, vol. 37, pp. 2027-2037, 2004.

[14] I. Hideyuki, T. Akira, and M. Masaaki, "A method of identifying influential data in fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol. 6, no. 1, 1998.

[15] R. W. Tibshirani, and G. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," *JRSSB*, 2000.

[16] G. Hamerly and C. Elkan, "Learning the k in k-means," NIPS, 2003.

[17] B. Jurgen and H. Eyke, "Adaptive optimization of the number of clusters in fuzzy clustering," in *Proc. 2007 IEEE International Fuzzy Systems Conference*, pp. 1-6.

[18] S. Michio and Y. Takahiro, "A fuzzy-logic-based approach to qualitative modeling," *IEEE Transaction On Fuzzy Systems*, vol. 1, no. 1, Feb. 1993.

[19] L. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11.

Seyed Mahmood Hashemi was born in 1979 at Iran. He get his BSc and MSc degrees from Islamic Azad University. Now, He is a PhD candidate in Beijing University of Technology (BJUT).