Using Fuzzy Rules in Identifying Cybercrime in Iranian Banking System

Omran Alipour, Saeid Khalilpour

***Abstract*— Similar growth of security and information technology and non-slot between these two subjects are factors of comfort in human societies. Therefore, base on evidences, with popularity and prevalence of using internet, cybercrime increases everyday because of failure of achieving balanced growth points, so that methods of attacks and fraud have become more complex. Therefore, security of cyberspace is major concern of banks, corporations and insurance companies. The main goal of this paper is using fuzzy algorithm and recommending effective system's cybercrime identification. Proposed frameworks, identifies and reports suspected cases in two levels. First level is analysis of user information and second one is detecting wrong warnings.**

***Index Terms*— FCM, decision tree, clustering, classification, cybercrime.**

I. INTRODUCTION

Because of the extent and complexity of information (big data), using tools for data mining and knowledge extraction is important and cybercrimes can be identified and detected in different areas like violence, exploitation and showing obscenity base on information technology capacity of analyze by detecting relations between evidences base on fuzzy rules that are closer to natural life.

Data mining process and knowledge extraction contain clustering, decision tree etc. methods to reveal hidden patterns. The required data mining should identify fraudulent and deceptive transactions by using clustering and classification techniques.

The main goal of this technique, is to identify cybercrimes in internet banking. Other parts of this paper has organized as following. The Second part describes main idea of paper; the third part introduces the suggested method and the fourth part presents general discussion and results.

II. THE BASIC IDEA

In the recommended framework, by using FCM clustering algorithm, two normal and suspected clusters is resulted base on paper authors banking experiences and then suspected cluster by C4.5 algorithm are classified to (warning, crime, no crime) class. In the other words, in this method combination of two algorithms has proposed for reducing wrong positive. Therefore, in the first level, needed calculations has done on general data (GD) and in second level, classification of suspected cases happens base on cybercrime's available data (CD).

III. BACKGROUND

A. CLASSIFICATION

With an abstract look to classification of data topic, it is possible to reach a two-leveled process. In first level, we made a model that determines a set of data classes or concepts. In this level, we made a model by analyzing an instructional set (selected set of tuples in database) that determines related classes to these tuples. A tuple of X is shown as X=(x1, x2, …, xn). Assume that every tuple belongs to a pre-described class and the class is identified by an adjective which we call it class sticker adjective. In this paper, instructional set is selected randomly from cybercrime database (CD).

In second level, learning happens by the function y=f(x) that can predict sticker class of every X tuple from base. This function is as classification, decision trees or mathematics formula rules. What we study here for classification is classifying by decision trees.

B. DECISION TREES

Decision tree is a unique method of a system that facilitates decision making in future and describes system properly. Properties and outputs can be analyzed by using an algorithm (tree making), and the system can be presented in the form of a tree; because most of engineering, administrative and computing systems can be defined in the form of a set of data (property or properties and matching output with them).

After making tree, it can be used created structure for classifying tests. In this method, test data made decisions in different classes for isolation by the fast structure.

Generally, training data that are used for tree making differ from test data that is used for tree evaluation and error number in determining test data classes is a criterion that shows the algorithm is suitable.

C. C5, C4, 5, ID3

ID3 algorithm has introduced in 1986 by Quinlan. In this algorithm it has been used the concept of irregularities for grouping data and algorithm tries to decrease the irregularities in upper part of tree to have a tree with the least height. So that, irregularity is calculated by formula

Omran Alipour, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran,
Saeid Khalilpour, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

1 for all primary data properties at first and then the property that contains the most advantage is selected as root.

$$I = -\sum_{c} p(c) \log_2 p(c) \qquad (۱)$$

Formula 2 used for calculating the advantage of every property.

$$Gain(A) = I - I_{res}(A) \qquad (۲)$$

(res) is the remaining irregularity rate in groups through the use of a property that can be obtained with the help of total probability of each of the divisions (formula 3).

$$I_{res} = -\sum_{v} p(v) \sum_{c} p(c|v) \log_2 p(c|v) \quad (۳)$$

This algorithm just has the ability of grouping data with the range of discrete and limited features and does not work in noisy and distorted data.

As we said before C4.5 algorithm is completed form of ID3. This algorithm has the ability of grouping noisy and continuous data. Therefore, we classify data at first, then, we make all situations that has the possibility of data separation and choose corresponding isolator with the maximum amount of advantage as isolator.

The last version of decision tree C4.5 is commercial and it has called C5. This version has the capability of analyzing multi-million records and uses multi-core CPU processing capabilities to make algorithm functionality better.

D. CLUSTERING METHOD FCM

In pattern identifying, cluster is a set of data that placed in a group because of their similarity to each other. In clustering it is tried to share set of data that is without controller to clusters to maximize the similarity of inner data of every cluster and minimize the similarity of inner clusters data. [2] One of the common ways of clustering methods of clustering is FCM method. FCM clustering method is part of the family of clustering algorithms that have **goal function** and their goal is minimizing a function.

In this algorithm the samples share into cluster c, which has a pre-determined number, and in c algorithm clustering fuzzy mean, goal function is as formula 4:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \; ||x_k - v_i||^2 \qquad (۴)$$

In formula 4, m is a real number that is bigger than 1 which is considered 2 most of the times. If we place 1 in front of m in the above formula we will have non-fuzzy mean of c clustering goal function. In the above formula xk is k sample, vi is center of cluster and n is number of samples. Uik shows the rate of I sample obtaining in k cluster. ||*|| sign is similarity rate of sample with cluster center that can be used any function that shows sample similarity and cluster center. It can be defined a U matrix with regard to uik which contains c lines and n columns and its components can use any amount between 0 and 1. If all of the components of U matrix is 0 or 1 c similar algorithm will be classic mean. Components of U matrix can choose any amount between 0 and 1, but total component of every column should be 1:

$$\sum_{i=1}^{c} \mathrm{u_{ik}} = 1, \forall \mathrm{k} = 1, \dots, n \qquad (۵)$$

Meaning of this condition is that total obtaining of every sample to cluster c should be 1. For having related formula to vi and uik we should minimize the defined goal function. By using the above condition and considering derivative function 0 the following equation is resulted:

$$u_{\mathrm{ik}} = \frac{1}{\sum_{\mathrm{j=1}}^{\mathrm{c}} (\frac{\mathrm{d_{ik}}}{\mathrm{d_{jk}}})^{2/(\mathrm{m}-1)}} \qquad (۶)$$

$$v_{\mathrm{i}} = \frac{\sum_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{u_{ik}^{m} x_k}}{\sum_{\mathrm{k=1}}^{\mathrm{n}} \mathrm{u_{ik}^{m}}} \qquad (۷)$$

By using formula 6 and 7, c clustering algorithm of fuzzy mean is as following.

Algorithm levels:

1- First amounting for c, m and U it must be guessed.
2- Cluster centers should be calculated (calculating vi)
3- Calculating obtaining matrix base on calculated clusters in 2
4- If it is as this equation ||Ui+1-Ui|| ≤ ε algorithm finishes and vice versa it goes to level 2.

Strength of fuzzy mean c algorithm is always being covergant and non-controller. It can be pointed from weak points of fuzzy mean of c algorithm to maximum time of calculation time and sensitivity to the first guesses and noise and it is possible for algorithm to stop in local minimums.

Sensitivity of algorithm to noise, has been soluted by considering a cluster for noisy data. Meanwhile, primary guesses for cluster center is recovered when some labeled data are available.

E. FUZZY RULES BASE

Fuzzy logic proposed by professor Zade for the first time and it has been used in many scienes and industries up to now. The first and the most successful usage of fuzzy logic are in control and the most important aspect of this theory is implementation of complex systems in the form of variables and fuzzy systems. Variables that is defined as fuzzy (continous and as an interval between 0 and 1) and the rules that are match with human knoeledge and are closer to objective usage.

Fuzzy rules base contains the rules that apply as IF-THEN as fuzzy variables. The way “and” or “or” has been defined base on expert diagnose and “aggregation” and “implication” should select correctly too. In this paper, the goal is proposing a way for producing rules and in the next part we will describe it completely.

F. BOOTSTRAP TECHNIQUES

Bootstrapping is a statistical-calculation method for determining the result of data sample estimator accuracy. In this technique we can just estimate by a simple method any statistic of sample data distribution. Generally, this method is one of the open methods of sampling.

Bootstrapping is estimating properties of the whole estimator by using the measuring of these properties in an approximate distribution. Bootstrapping can be implemented by making some sample that every one of these samples are random or replaced samples of the whole main data.

Bootstrapping can be use for hypothesis statistical test too. This method is using generally as an alternative for inferential method base on parametric hypothesis when we suspect on these hypothesis.

Bootstrap application:

1- When the statistic distribution of is complex or unknown
2- When size of a sample is not enough for statistical inference

IV. RECOMMENDED FRAMEWORK

Recommended framework’s levels are as followings:

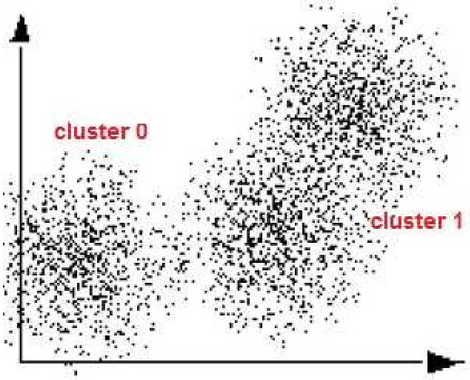
Clustering levels:

1) Beginning
2) Selecting a set of “Data warehouse” tuples base on bootstrap technique (S)
3) Performing FCM algorithm for S for any properties of A1 to A4
4) Finding appropriate threshold
5) Two normal and suspected cluster will be result of this threshold
6) Selecting a set of CD base on bootstrap technique
7) The end

The suspected cluster will propose as c4.5 classification of algorithm input as the output of this level.

Classification

1) Beginning
2) Selecting set of CD data base tuples base on bootstrap technique
3) Making decision tree base on C4.5 algorithm and A1 to A4 properties
4) Classifying the suspected cluster
5) The end



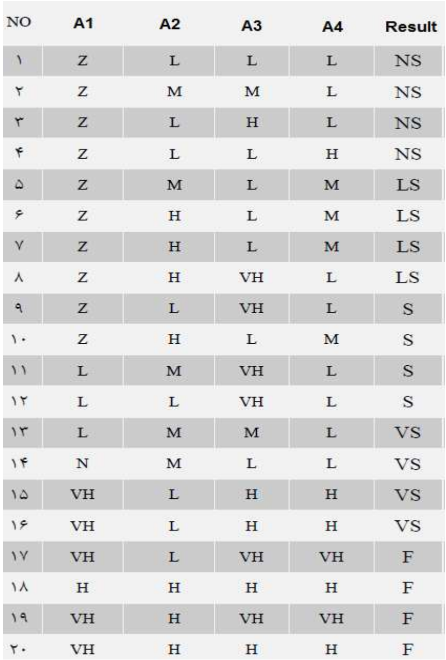
Picture 1

In this method decision tree has made by analyzing input data in every node and studying every property. Basically, using decision tree is for making rules which is effective in results by using FCM in data isolation. For clarifying proposed method, consider some data as an example that contains four properties and are classified as three classifications. Selecting a property in every node means that this property has isolated data more than other ones, therefore it helps us to detect the crime as result in the final tree. C4.5 algorithm selects data in a way that chooses the best property for minimizing irregularity by using entropy concept and a threshold is going to select for sharing data and their decision.

In order to prevention of noise by bootstrap technique, we choose suitable data from GD database and entrust it to FCM algorithm. The result will be two clusters 0 and 1 that are representative of two normal and suspected clusters. At the same time, we make decision tree base on C4.5 algorithm and CD database tuples that is because of the following variables which is base on expert opinion. In the following tree we will have 3 warning, crime and no crime leaves.

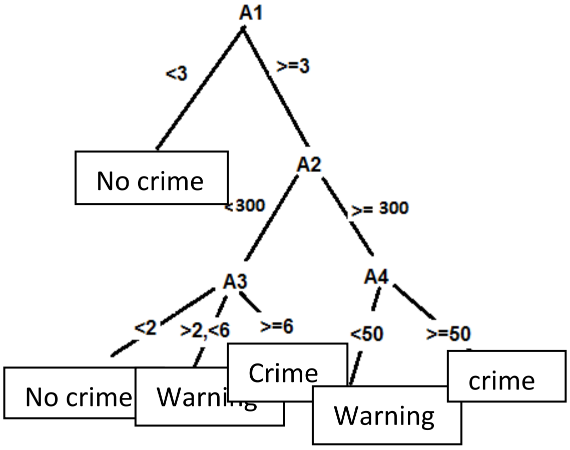
A1	Number of unsuccessful attempts of user
A2	Continuous and non-stop activity of user
A3	Daily debtor documents as milliard
A4	Unconventional check of account balances

Table 1



NO	A1	A2	A3	A4	Result
١	Z	L	L	L	NS
٢	Z	M	M	L	NS
٣	Z	L	H	L	NS
۴	Z	L	L	H	NS
۵	Z	M	L	M	LS
۶	Z	H	L	M	LS
٧	Z	H	L	M	LS
٨	Z	H	VH	L	LS
٩	Z	L	VH	L	S
١٠	Z	H	L	M	S
١١	L	M	VH	L	S
١٢	L	L	VH	L	S
١٣	L	M	M	L	VS
١۴	N	M	L	L	VS
١۵	VH	L	H	H	VS
١۶	VH	L	H	H	VS
١٧	VH	L	VH	VH	F
١٨	H	H	H	H	F
١٩	VH	H	VH	VH	F
٢٠	VH	H	H	H	F

Table 2- expert opinion



Picture 2

Adjective vectors in cluster 1 (suspected) applied to the above decision tree. It is necessary to notice to corresponding tuples of adjective vectors that lead to crime and warning leaves and start fieldwork from these users.

V. CONCLUSION

Along with the complexity of methods of fraud should move toward developing defensive techniques by using machine learning. It should have used FCM algorithm for isolating data and make the rules by high performance. The rules will extract with related variables and outputs of system is studying under different situation and with regard to suitable performance of the constructed decision tree, the rules are evaluated and the impeachment results will propose with the least error which has high advantages like analyzing behavior of an organization's users and detecting organizational risks.

REFERENCES

[1] R. kosari, N. Moqadam, D. Vahdat. 1392. Using algorithm of decision tree to detect suspected behaviors in internet banking. Tehran: Research Quarterly Journal of Science and Technology Information Institute of Iran

[2] M. Yaqmaie, 1388. Introducing new structure of decision tree and using it along with clustering (FCM) for producing fuzzy rules, the fifteenth conference of computer society of Iran

[3] Quinlan, J. R. Induction of Decision Trees. *Mach.Learn.*, 1986

[4] Adaptive Algorithm for Cyber Crime Detection. Manveer Kaur, Sheveta Vashisht, Kumar Saurabh. 2012

[5] Ji Dan, Qiu Jianlin, Gu Xiang, Chen Li, He Peng, A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree", 10th IEEE International Conference on Computer and Information Technology .2010

[6] Alex Berson, Stephen J. Smith, "Data Warehousing, Data Mining, & Olap" by Tata McGraw-Hill. Teknomo, Kardi ," K-Means Clustering Tutorials".

[7] http://databases.about.com/od/datamining/a/kmeans.htm

[8] http://en.wikipedia.org/wiki/C4.5_algorithm

[9] http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

[10] JohannesGehrke_,VenkateshGanti, Raghu Ramakrishnany and Wei-Yin Lohz,"BOAT— Optimistic Construction", Proceedings of the 1999

[11] ACM SIGMOD Decision Tree International conference on Management of data, 1999

[12] Manveer Kaur et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, 2012,