Vol. 8(27), Jan. 2018, PP. 3764-3771

Contents list available at IJMEC

International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)

Journal Homepage: www.aeuso.org

PISSN: 2411-6173 - EISSN: 2305-0543

Fraud Detection in Automobile Insurance using a Data Mining Based Approach

Ali Ghorbani and Sara Farzai*

[1] *Department of Industrial Engineering, Faculty of Engineering, Maziar University, Noor, Iran*

Phone Number: *+98-1144904*

****Corresponding Author's E-mail:*** *s.farzai@maziar.ac.ir*

Author's E-mail: *a.ghorbani@maziar.ac.ir*

Abstract

Insurance industry is one of the most important issues in both economy and human being life in modern societies which awards peace and safety to the people by compensating the financial risk of detriments and losses. This industry, like others, requires to choose some strategies to obtain desired ranking and remain in competitive market. One of efficient factors which affects enormous decision makings in insurance is paying attention to important information of customers and bazar that each insurance company stores it in its own database. But with daily increasing data in databases, although hidden knowledge and pattern discovery using usual statistical methods is not impossible, it is so complicated and time-consuming. In this paper we employ data mining as a powerful approach for extracting hidden knowledge and patterns on massive data to guide insurance industry. For example, one of the greatest deleterious challenges here is interacting between insurance companies and policyholders which creates a feasible situation for fraudulent claims. Due to importance of this issue, after investigating different ways of fraudulent crimes in insurance, we use K-Means clustering technique to find fraud patterns in automobile insurance include body and third-party. Our experimental results indicate a high accuracy when have been compared with statistical information extracted from data sets. Outcomes show significant relations among efficient factors in similar fraud cases.

Keywords: *automobile insurance, data mining, clustering, k-means.*

1. Introduction

Nowadays that a variety of disasters threatens people's life and wealth, insurance is a desirable choice to carry these risks from insured people to insurance companies. Insurance is a contract between the insurer and the insured that implies if losses which are detailed in policy occur, policyholder will be financially compensated by insurer up to a predetermined bound instead of insurance premium that has been paid. Insurance companies present a variety of services each of which can cover a part of losses. Beside fast development of information technology, the amount of stored data in insurance companies' databases is growing rapidly. These vast databases contain essentially useful commercial information. In the other hand, discovering worthwhile hidden information in these databases and also identifying suitable models are not so straightforward. One

Article History:

Received Date: Jul. 09, 2017
Accepted Date: Oct. 23, 2017
Available Online: Jan. 05, 2018

of the efficient methods for discovering hidden knowledge and finding patterns on big databases is data mining. In recent two decades many researches are presented around fraud in automobile insurance. Belhadji and Dionne (1997) in their research investigated on key factors about insurance fraud. They consulted with domain experts in this area. They calculated conditional probability of fraud for each factor and then they determined the most important factors and also fraudulent losses by using regression algorithm. Cummins and Tennyson (1992) studied on a similar topic. They firstly refined the costs of insurance services and the rate of cost increasing with corresponding service improvement.

Then they detect efficient factors in insurance inflation and specially the factors that grow with cost increasing. In the cases with invulnerable rules and guaranteeing the ease of debt payback, their proposed method was efficient. Brockett et al. (1998) classified the automobile body insurance frauds using Kohenen self-organizing tool, and with considering the level of fraud suspicion. At first they selected specifications using Principal Component Analysis (PCA) algorithm. Then they detect body insurance frauds by a combination of clustering and back propagation artificial neural networks. Weisberg and Derrig (1993) proposed some techniques for detecting fraudulent losses and frauds classification.

Their research illustrated a supposition of fraud. Also they proposed definitions of fraudulent crimes and quantitative values. Derrig and Ostazewski (1995) clustered risks and classified frauds with an overview on fuzzy sets and fuzzy pattern detection techniques. This research continued the approach of Weisberg and Derrig (1993). In other research presented by Artis et al. (2002), a method for modifying the kind of losses was proposed.

They compared Multinomial Logit Model (MLM) and Nested Logit Model (NLM) models in detecting automobile insurance fraud. Viaene and Dedene (2004) presented an approach based on simple Bayes to detect fraud in automobile insurance data. They also used boost algorithms in their research and compared the results. Their results indicated that boost algorithms lead to more exact answers. Phua et al. (2004) combined back-propagation neural networks, simple Bayes and decision tree to detect fraud in automobile insurance. The result showed that proposed method produces better answers in comparison with the best prior classifiers, from time-saving point of view.

In this research, we use K-Means clustering algorithm and Weka software to detect fraud in automobile insurance specifically body and third-party with investigating on 100 samples extracted from insurance companies. The goal of this work is to detect the possibility of fraud in similar cases according to common specifications with our fraudulent cases.

The rest of the paper is arranged as follows: in section 2 we briefly define required literature and preliminaries, section 3 represents our clustering based method, experimental results are shown in section 4, and section 5 is dedicated to conclusion.

2. Literature overview

Before explaining our clustering method in section 3, keywords and expressions are required to be defined briefly as is came in the following.

2.1. Insurance

Insurance is an agreement where, for a stipulated payment called the premium, one party (the insurer) agrees to pay to the other (the policyholder or his designated beneficiary) a defined amount (the claim payment or benefit) upon the occurrence of a specific loss. This defined claim payment amount can be a fixed amount or can reimburse all or a part of the loss that occurred. Each premium may be adjusted to reflect any special characteristics of the particular policy [1].

Under the formal arrangement, the party agreeing to make the claim payments is the insurance company or the insurer. The pool participant is the policyholder. The payments that the policyholder makes to the insurer are premiums. The insurance contract is the policy. The risk of any unanticipated losses is transferred from the policyholder to the insurer who has the right to specify the rules and conditions for participating in the insurance pool [1].

The insurer may restrict the particular kinds of losses covered. For example, a peril is a potential cause of a loss. Perils may include fires, hurricanes, theft, and heart attack. The insurance policy may define specific perils that are covered, or it may cover all perils with certain named exclusions (for example, loss as a result of war or loss of life due to suicide) [1].

2.2. Insurance Fraud

Insurance fraud is any act committed with the intent to obtain a fraudulent outcome from an insurance process. This may occur when a claimant attempts to obtain some benefit or advantage to which they are not otherwise entitled, or when an insurer knowingly denies some benefit that is due. According to the United States Federal Bureau of Investigation the most common schemes include: Premium Diversion, Fee Churning, Asset Diversion, and Workers Compensation Fraud. The perpetrators in these schemes can be both insurance company employees and claimants [2]. False insurance claims are insurance claims filed with the intent to defraud an insurance provider.

Insurance fraud has existed since the beginning of insurance as a commercial enterprise [3]. Fraudulent claims account for a significant portion of all claims received by insurers, and cost billions of dollars annually. Types of insurance fraud are diverse, and occur in all areas of insurance. Insurance crimes also range in severity, from slightly exaggerating claims to deliberately causing accidents or damage. Fraudulent activities affect the lives of innocent people, both directly through accidental or intentional injury or damage, and indirectly as these crimes cause insurance premiums to be higher. Insurance fraud poses a significant problem, and governments and other organizations make efforts to deter such activities.

2.3. Data mining and clustering

Data mining is an evolution in database systems and database applications. Data mining, also called Knowledge Discovery in Databases (KDD), is the skill of extracting patterns which represent the knowledge implicitly stored in large databases, data warehouses, and other massive information repositories.

Data mining is a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, knowledge acquisition, information retrieval, high performance computing, and data visualization.

Data mining emerged during the late 1980's, has made great strides during the 1990's, and is expected to going-on into the new millennium.

Simply stated, data mining refers to extracting or “mining" knowledge from large amounts of data. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material [4].

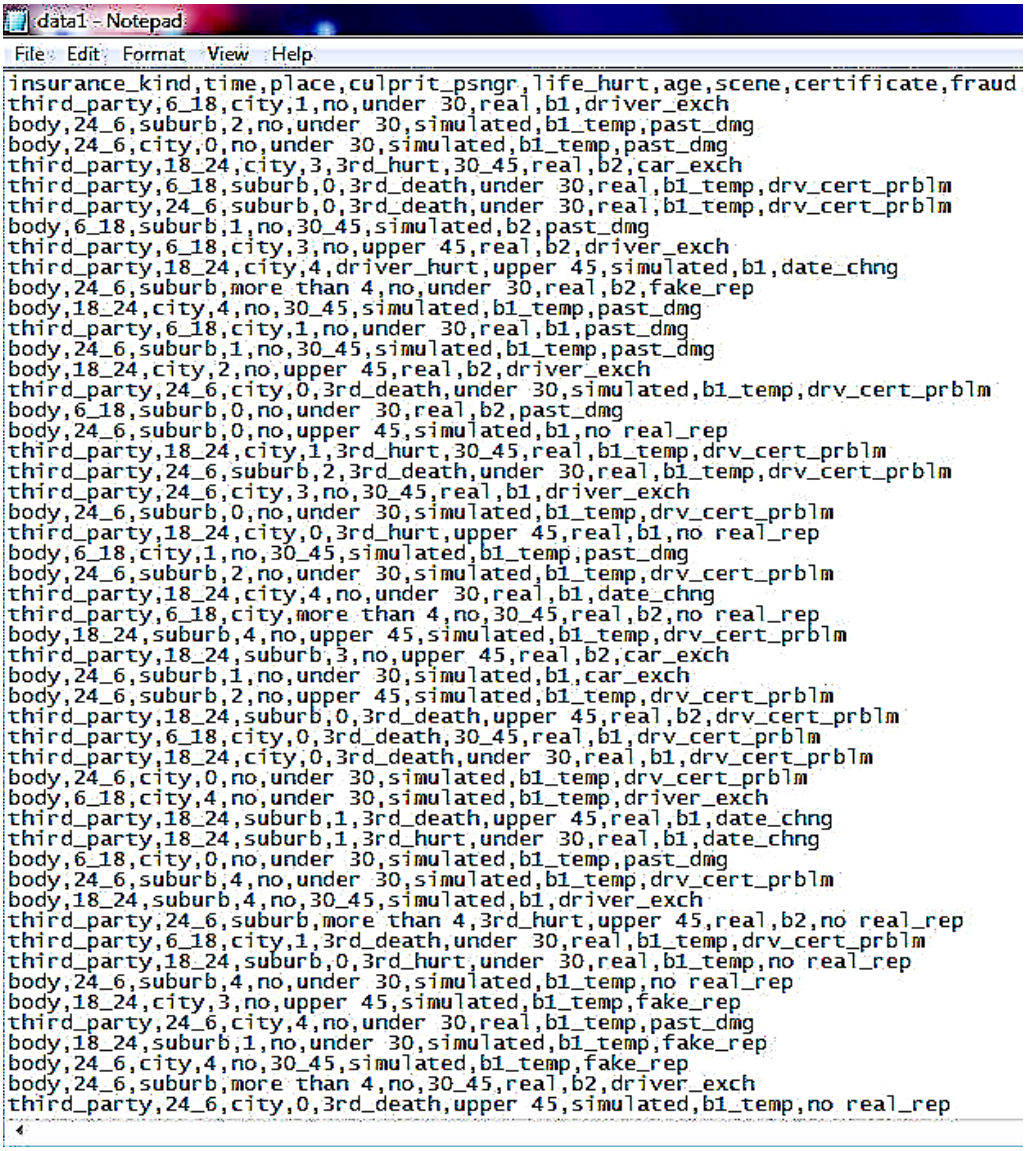
Among the most used data mining techniques, classification analyzes class-labeled data objects in training step, and creates a model which can predict other similar objects. Whereas clustering analyzes data objects without consulting any known class label.

Generally in clustering, the class labels are not determined in the training data. Because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity, that is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Each cluster can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together. K-Means is a simple clustering method which is the basis of other more complicated clustering methods, and is our preferred technique in this paper that we explain it in following [4].

3. Fraud detection using K-Means clustering

Deviation and fraud would be found in most of the economic activities. More roles in inspecting a payment case lead to more risks of fraud. One of the ways of detecting fraud in reported damages and losses is to use information remained from former detected fraudulent cases. In this section we present a method based on K-Means clustering to detect fraud patterns in automobile body insurance and third-party insurance.


data1 - Notepad

File Edit Format View Help

```
insurance_kind,time,place,culprit_psngr,life_hurt,age,scene,certificate,fraud
third_party,6_18,city,1,no,under 30,real,b1,driver_exch
body,24_6,suburb,2,no,under 30,simulated,b1_temp,past_dmg
body,24_6,city,0,no,under 30,simulated,b1_temp,past_dmg
third_party,18_24,city,3,3rd_hurt,30_45,real,b2,car_exch
third_party,6_18,suburb,0,3rd_death,under 30,real,b1_temp,drv_cert_prblm
third_party,24_6,suburb,0,3rd_death,under 30,real,b1_temp,drv_cert_prblm
body,6_18,suburb,1,no,30_45,simulated,b2,past_dmg
third_party,6_18,city,3,no,upper 45,real,b2,driver_exch
third_party,18_24,city,4,driver_hurt,upper 45,simulated,b1,date_chng
body,24_6,suburb,more than 4,no,under 30,real,b2,fake_rep
body,18_24,city,4,no,30_45,simulated,b1_temp,past_dmg
third_party,6_18,city,1,no,under 30,real,b1,past_dmg
body,24_6,suburb,1,no,30_45,simulated,b1_temp,past_dmg
body,18_24,city,2,no,upper 45,real,b2,driver_exch
third_party,24_6,city,0,3rd_death,under 30,simulated,b1_temp,drv_cert_prblm
body,6_18,suburb,0,no,under 30,real,b2,past_dmg
body,24_6,suburb,0,no,upper 45,simulated,b1,no real_rep
third_party,18_24,city,1,3rd_hurt,30_45,real,b1_temp,drv_cert_prblm
third_party,24_6,suburb,2,3rd_death,under 30,real,b1_temp,drv_cert_prblm
third_party,24_6,city,3,no,30_45,real,b1,driver_exch
body,24_6,suburb,0,no,under 30,simulated,b1_temp,drv_cert_prblm
third_party,18_24,city,0,3rd_hurt,upper 45,real,b1,no real_rep
body,6_18,city,1,no,30_45,simulated,b1_temp,past_dmg
body,24_6,suburb,2,no,under 30,simulated,b1_temp,drv_cert_prblm
third_party,18_24,city,4,no,under 30,real,b1,date_chng
third_party,6_18,city,more than 4,no,30_45,real,b2,no real_rep
body,18_24,suburb,4,no,upper 45,simulated,b1_temp,drv_cert_prblm
third_party,18_24,suburb,3,no,upper 45,real,b2,car_exch
body,24_6,suburb,1,no,under 30,simulated,b1,car_exch
body,24_6,suburb,2,no,upper 45,simulated,b1_temp,drv_cert_prblm
third_party,18_24,suburb,0,3rd_death,upper 45,real,b2,drv_cert_prblm
third_party,6_18,city,0,3rd_death,30_45,real,b1,drv_cert_prblm
third_party,18_24,city,0,3rd_death,under 30,real,b1,drv_cert_prblm
body,24_6,city,0,no,under 30,simulated,b1_temp,drv_cert_prblm
body,6_18,city,4,no,under 30,simulated,b1_temp,driver_exch
third_party,18_24,suburb,1,3rd_death,upper 45,real,b1,date_chng
third_party,18_24,suburb,1,3rd_hurt,under 30,real,b1,date_chng
body,6_18,city,0,no,under 30,simulated,b1_temp,past_dmg
body,24_6,suburb,4,no,under 30,simulated,b1_temp,drv_cert_prblm
body,18_24,suburb,4,no,30_45,simulated,b1,driver_exch
third_party,24_6,suburb,more than 4,3rd_hurt,upper 45,real,b2,no real_rep
third_party,6_18,city,1,3rd_death,under 30,real,b1_temp,drv_cert_prblm
third_party,18_24,suburb,0,3rd_hurt,under 30,real,b1_temp,no real_rep
body,24_6,suburb,4,no,under 30,simulated,b1_temp,no real_rep
body,18_24,city,3,no,upper 45,simulated,b1_temp,fake_rep
third_party,24_6,city,4,no,under 30,real,b1_temp,past_dmg
body,18_24,suburb,1,no,under 30,simulated,b1_temp,fake_rep
body,24_6,city,4,no,30_45,simulated,b1_temp,fake_rep
body,24_6,suburb,more than 4,no,30_45,real,b2,driver_exch
third_party,24_6,city,0,3rd_death,upper 45,simulated,b1_temp,no real_rep
```

Figure 1: A part of CSV file to be used in Weka

Our statistical population consists of 100 fraudulent cases gathered from Iranian insurance companies: Iran, Asia, Alborz, Dana, Dey and Pasargad. Then we choose a number of most common ways to simulate losses and damages in a fake manner, as following:

1. insurance type: third-party / body
2. time of occurrence: 6-18 / 18-24 / 24-6
3. position of occurrence: urban / suburbia
4. number of passengers in guilty car (driver is not considered): 0 / 1 / 2 / 3 / 4 / more than 4
5. accident led to injury: no injury / guilty driver injury / guilty driver passing away / third-party injury / third-party passing away
6. age of driver: less than 30 / between 30 and 45 / more than 45
7. scene of accident: real / recreated
8. license type: B-1(temporary) / B-1 / B-2
9. fraud type: fake accident report / fake croquis / reporting older damages in current croquis / changing date of accident / replacing the driver who has license problem / exchanging guilty driver and lost driver / exchanging guilty car and damaged car.

From statistical point of view, up to 33% of all the 100 fraudulent cases are related to drivers that there are some problems in their licenses, such as expired licenses, or a driver with B-1(temporary) license who has an accident in suburbia. Reporting older losses and damages in current accident with 19% is in next place, and fake accident reports and fake croquis stand at third place with 11%. These statistics will be used to evaluate precision of the patterns detected by our method in comparison to reality.

The mentioned information has been gathered by consulting domain experts at insurance companies. Then we transform it into CSV file format using Microsoft Excel program, to be used in Weka software for clustering process. A part of CSV file is shown in figure 1.

We choose 7 clusters to grouping data points according to expert suggestion, and Euclidean distance is used to measure similarity and dissimilarity among data points and clusters. Figure 2 illustrates the results of our K-Means clustering process, and visual results based on type of fraud and type of injury are shown in figure 3.

```
Cluster centroids:
                                    Cluster#
Attribute               Full Data          0          1          2          3          4          5          6
                          (100)         (14)       (27)       (12)       (17)       (16)        (8)        (6)
=============================================================================================================
insurance_kind       third_party third_party       body       body third_party       body third_party third_party
time                       24_6       18_24       24_6       24_6       18_24       18_24       6_18       24_6
place                   suburb        city       city     suburb       city     suburb       city     suburb
culprit_psngr              0.0 more than 4        4.0        0.0        0.0        1.0        1.0        2.0
life_hurt                   no          no         no         no   3rd_hurt         no         no  3rd_death
age                    under 30       30_45   under 30   under 30   upper 45   under 30   under 30   under 30
scene                      real        real  simulated       real       real  simulated       real       real
certificate             b1_temp          b2    b1_temp         b1         b1    b1_temp         b1    b1_temp
fraud           drv_cert_prblm driver_exch drv_cert_prblm past_dmg drv_cert_prblm drv_cert_prblm drv_cert_prblm drv_cert_prblm



Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       14 ( 14%)
1       27 ( 27%)
2       12 ( 12%)
3       17 ( 17%)
4       16 ( 16%)
5        8 (  8%)
6        6 (  6%)
```

Figure 2: The results of K-Means clustering process.

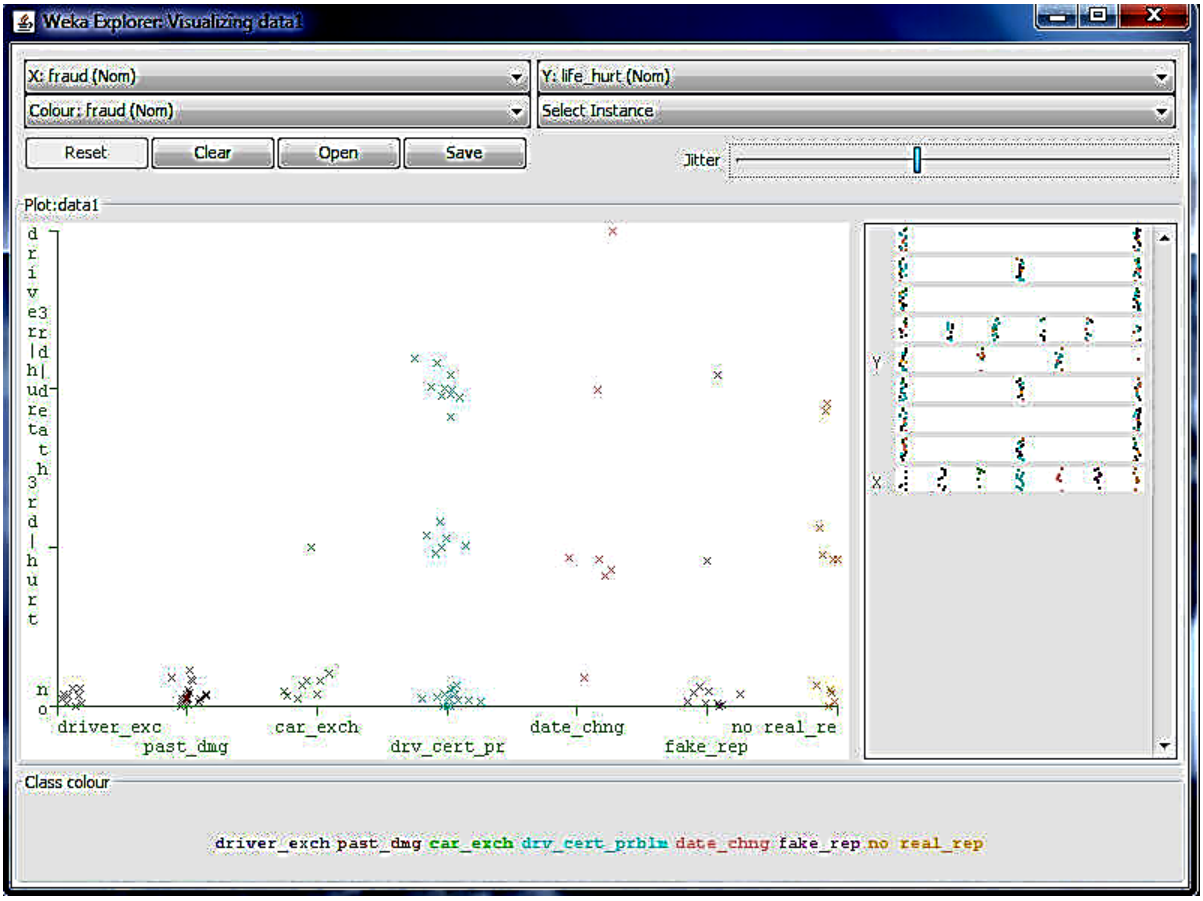


Figure 3: Visual results based on type of fraud and type of injury.

As is shown in figure 3, most compression of data points could be seen in clusters which injuries are not contained in them. In the cases of injuries, such as driver (third-party) injury or passing away, the most detected frauds are in relation with the licenses of guilty drivers. We evaluate these outcomes in comparison to real statistics in next section.

4. Comparison and evaluation

From figure 2 we can extract some patterns and propositions as below:

- **in cluster #0,** 14% of all the 100 data points have similarity in these items: the accident occurred at the first hours of night, in urban territories, with no injury, B-2 license, fraud in third-party insurance with type of exchanging guilty driver and lost driver.

- **in cluster #1,** 27% of all the 100 data points have similarity in these items: the accident occurred at the hours of midnight, in urban territories, with no injury, driver was younger than 30, accident scene is recreated, B-1 license(temporary), fraud in body insurance with type of license problem.

- **in cluster #2,** 12% of all the 100 data points have similarity in these items: the accident occurred at the hours of midnight, in suburbia territories, with no injury, driver was younger than 30, B-1 license, fraud in body insurance with type of reporting older damages in current croquis.

- **in cluster #3,** 17% of all the 100 data points have similarity in these items: the accident occurred at the first hours of night, in urban territories, led to third-party injury, driver was older than 45, B-1 license, fraud in third-party insurance with type of problem in license of guilty driver.

- **in cluster #4,** 16% of all the 100 data points have similarity in these items: the accident occurred at the first hours of night, in suburbia territories, with no injury, driver was younger than 30, B-1 license(temporary), fraud in body insurance with type of problem in license of driver.

- **in cluster #5,** 8% of all the 100 data points have similarity in these items: the accident occurred at the hours of day long, in urban territories, with no injury, driver was younger than 30, B-1 license, fraud in third-party insurance with type of problem in license of guilty driver.

- **in cluster #6,** 6% of all the 100 data points have similarity in these items: the accident occurred at the hours of midnight, led to third-party passing away, driver was younger than 30, B-1 license(temporary), fraud in third-party insurance with type of problem in license of guilty driver.

Generally, most of the detected fraudulent cases have been occurred at first hours of night or hours of midnight, with no injury, that driver was younger than 30 and had B-1(temporary) license; and the most of the frauds are seen in body insurance with type of problem in license of driver.

Outcomes are too close to real statistics. It indicates that in new accidents with similar positions and situations to our data set, with considering suspicion of new cases according to detected patterns, exact investigating about fraud possibility is required.

Conclusion

In this paper we used data mining and K-Means clustering technique to detect fraud in automobile insurance, third-party and body, and we found some patterns on fraudulent cases which can be helpful to detect new frauds. We chose 100 cases of automobile insurance fraud from Iranian insurance companies and then we refined the common properties among them. After gathering required information, we select 7 clusters according to insurance expert suggestion and K-Means clustering technique has been performed on our fraud data set with Euclidean distance as similarity and dissimilarity measurement. The results of proposed technique indicates a significant accuracy in

comparison to real statistics, meanwhile the proposed method succeed to extract some patterns and propositions which would be helpful to detect fraud in next accident cases. As future work, accuracy of fuzzy techniques can be studied to cover natural uncertainty of encountered events. Also the impact of some other properties such as geographic position, and insurance market basket can be investigated in automobile fraud occurrence.

References

[1] J. F. Anderson and R. L. Brown, "Risk and insurance", Education and examination committee of the society of actuaries, 2005, P-21-05.

[2] "FBI — Insurance Fraud". Fbi.gov, 2005-09-08, Retrieved 2014-02-07.

[3] A. Manes, "Insurance Crimes", p. 34.

[4] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", Third Edition Elsevier Inc, 2012.

[5] D. B. Belhadji and G. Dionne, "development of an expert system for the automatic detection of automobile insurance fraud", Risk Management Chair, HEC-Montreal, 1997.

[6] J. D. Cummins and S. Tennyson, "Controlling automobile insurance costs", Journal of Economic Perspectives, 1992, pp. 95-115.

[7] P. L. Brockett, X. Xia and R. A. Derrig, "Using kohonen's selforganizing feature map to uncover automobile bodily injury claims fraud", The J. of Risk and Insurance, 1998, pp. 245-74.

[8] H. I. Weisberg and R. A. Derrig, "Quantitative methods for detecting fraudulent automobile bodily insurance claims", AIB Cost Containment/Fraud Filing, 1993, pp. 49-82.

[9] R. A. Derrig and K. M. Ostazewski, "Fuzzy techniques of pattern recognition in risk and claim classification", The J. of Risk and Insurance, 1995, pp. 447-82.

[10] M. Artis, M. Ayuso and M. Guillen, "Detection of automobile insurance fraud with discrete choice models and misclassified claims", Journal of Risk and Insurance, 2002, pp. 325-40.

[11] S. Viaene and G. Dedene, "Insurance fraud: Issues and challenges", Geneva Papers on Risk and Insurance Issues and Practice, vol. 29, 2004, pp. 313-33.

[12] C. Phua, D. Alahakoon and V. Lee, "Minority report in fraud detection: classification of skewed data", Sigkdd Explorations, vol. 6, no. 1, 2004, pp. 50-9.

[13] L. Šubelj, Š. Furlan and M. Bajec, "An expert system for detecting automobile insurance fraud using social network analysis", Expert Systems with Applications: An International Journal, vol. 38, 2010, pp. 1039-52.