ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(7), DOI: 10.17485/ijst/2016/v9i7/87846, February 2016* ISSN (Online) : 0974-5645

Using the Clustering Algorithms and Rule-based of Data Mining to Identify Affecting Factors in the Profit and Loss of Third Party Insurance, Insurance Company Auto

Faramarz Karamizadeh[1] and Seyed Ahad Zolfagharifar[2*]

[1]Department of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran; f.karamizadeh@gmail.com
[2]Kohgiluyeh and Boyer Ahmad Science and Research Branch, Islamic Azad University, Iran; sir.zolfaghari@gmail.com

Abstract

Background/Objectives: Insurance data analysis can be considered as a way of losses reduction by using data mining. It uses the machine learning, pattern recognition and data base theory for discovering the unknown knowledge. **Methods/Statistical Analysis:** In this paper, information of 2011, third party insurance of Iran insurance company auto has analyzed in Kohgiluyeh and Boyer Ahmad by using the data mining method. **Findings:** The results show that using clustering algorithms with acceptable clusters will be able to provide a model to identify affecting factors and to determine the effect of them in the profit and loss of auto third party insurance. **Applications/Improvements:** The algorithm of K-Means has formed the best clustering with 9 clusters that have relatively good quality. It means that has been able to maximize the distance between the cluster and minimize the within cluster distance.

Keywords: Clustering Algorithm, Data Mining, Insurance, Profit and Loss, Third Party

1. Introduction

Insurance data analysis can be considered as a way of reducing insurance companies' losses and data mining may lead to useful results. Data mining is unknown knowledge and laws discovery process and useful of mass data and data bases[1]. Data mining is a useful tool for exploring knowledge from large data[2]. Because the data mining tools predict process and future behavior by monitoring data base for hidden patterns, cause to make decisions based on knowledge and easily respond to the questions that earlier was very time consuming[3]. Using the data mining (with supervisor or without supervisor) can achieve to discover of the hidden rules in the data[4].

[5]Presented a methodology using clustering data mining methods decision tree for management of insurance customers. The decision tree results with 99.66% accuracy showed that the main cause of customer churn is lack of satisfaction with the performance of the insurance company, high insurance premiums and so on.

[6]Had a research on the identification of fraud in auto insurance by using data mining. The results show that the simple Bayes algorithm with accuracy of 90.28% then decision tree with an accuracy of 88.9% and finally logistic regression with accuracy of 86.1% have been able to recognize the false or fraudulent of damage claim.

[7]Have done a research based on the classification of the policy holders' risk of auto insurance by using data

* *Author for correspondence*

mining algorithms. The aim of researchers is classification of insurance policy holders due to the risk of receiving or not receiving compensation during the insurance period in insurance company. First, they collected customer profile data that record 13768 during the years 2009-2010 and after the necessary pre-processing, run opposite algorithms on them and compare their results[8–11]. Used techniques included 6 cases, including decision tree, neural networks, Bayesian networks, support vector machines, logistic regression and discriminant analysis. The best accuracy between these algorithms related to decision tree that with accuracy of 76.4% could detect high-risk or low-risk of a customer.

2. Parts of the Research

In this section, we will study the data and area of study. Then the different stages of research and used methods investigate and the results of each part will be explained.

2.1 Data and Area of Study

In this study, first collected third-party damage and issued insurance policy in 2011, (about 20 thousand records, that 1500 record had damage) that includes 179 fields on the issued data. Then 137 fields that were not effective were omitted and at the end, effective fields decrease to 42 fields. The insurance experts also were considered to reduce the scale of the problem for removing the various fields. From loss data set, just determined fields of the amount of damages and details are extracted. Unfortunately, there weren't the more useful information such as the age of the fault driver, education, etc. and because key information of issuing data use at the time of record damage for an insurance policy, given that the most important fields of issued data are available from the previous stage so, with the integration of damage fields and issued to a complete information will have access about a particular insurance policy (Tables 1 and 2).

Selected fields
The amount of damage
The date of accident creation
First injured insurer
The number of injured stricken
The number of deceased stricken

Operations of data mining was used by rapid miner software and to optimize the responses and quality of the results also have been used Minitab and Clementine 12 software.

Table 2. Selected fields of insurance policy issued data

Total records	Total effective records	Fields	
		Effective	Non_effective
20000	1500	137	42

Table 1. Selected fields of insurance policy loss data

Line	Field name	Line	Field name	line	Field name
1	Month	15	Surpluscommitment	29	start date
2	year	16	Physical commitment	30	Date of issue
3	Agency code of major exporter	17	Financial commitment	31	Organization Name
4	Group discounts	18	Insurance policy of the previous year	32	Policy Issues
5	Discount of no damage	19	Insurance	33	Employee
6	Type of Document1	20	Plaque	34	Issued by branch
7	Late penalty	21	More used	35	Government
8	Add code of premiumrate	22	Capacity	36	Representative of Issue place
9	premium	23	Number of cylinders	37	Damage?
10	29 Article complications	24	Year of construction	38	The amount of damage
11	Taxation	25	System	39	Date of accident creation
12	Seat premium	26	Type of vehicle	40	First injured insurer
13	Surplus premium	27	Term of insurance	41	The number of injured stricken
14	Legal party premium	28	Expiration date	42	The number of deceased stricken

Table 3. Statics of the third party insurance policy issued Kohgiluyeh and Boyer Ahmad in 2011

Field name	Number	Missing removal method
System	70	Diagnosis according to another features
Type of vehicle	33	Diagnosis according to another features
More used	11	Diagnosis according to another features
Number of cylinders	2	Diagnosis according to another features
Governmental	28	Diagnosis of the plaque
Month	130	Diagnosis of the issued date
Insurance	49	Diagnosis of the insurer name

2.2 The Steps of the Research

2.2.1 Investigate to Missing Data

In the initial phase attempts to discover the missing values with sorting all the features regularly in the Microsoft Excel software and through other characteristics of each record have guessed the missing amount. Also, lost amounts will be identified during data transfer to data mining area, including fields with missing values and trouble shooting methods are in Table 4.

Table 4. Fields with missing values and methods of trouble shooting

Type of field	Field name
Integer	Surplus commitment, Physical commitment, Financial commitment, Plaque, Capacity, Number of cylinders, Year of construction, Term of insurance, The number of injured stricken, The number of deceased stricken
Polynominal	More used, System, Type of vehicle, First injured insurer
Real	Late penalty, Add code of premium rate, premium, Taxation, Seat premium, Legal party premium, The amount of damage
Binominal	Insurance policy of the previous year, Employee, Issued by branch

The number of records that have had lost values in several important features and removed and have been about 350 cases.

2.2.2 Remote Data Discovery

For detection of outliers data, box plot graphs and Minitab15 software was used. In this graph, percentile concept use that data between 25% and 75% are shown respectively with Q1and Q3 the most important part of data. X50% also shows the median and is determined with a line in the middle of the graph. Inter Quartile Range (IQR) is another concept that is also IQR = Q3-Q1.

Values greater than Q3 + [(Q3 - Q1) X 1.5] and less than Q1 - [(Q3 - Q1) X 1.5] are outliers data. To do this, implement box plot graph on the individual characteristics of the data and carried away data were corrected according to the results.

2.2.3 The Type and Name of Data Fields

At this stage, given the knowledge that is derived from data fields, proceed to determine the type of data for the software. Selected fields are according to Table 5.

Table 5. Determination of type and field name

Rule	Result	Support	Confidence
No surplus commitment, add code of premium rate	Damage	32%	41%

2.3 Evaluation Criteria Rule-Based Algorithm (The Discovery of Association Rules)

Association rules produce many patterns that may not be attractive all models for us. So, the criteria should be defined for assessing the quality of the rules. If you have a rule that says A, then B, of the number of records where A, B are both present, the total number of records, a measure is obtained that named Support. The numerical value is between 0 and 1. Usually in search of better rules, consider a threshold for support to be limited the number of obtained rules.

Threshold value may be cause to not see the rules that their support is lower than the threshold but also be valuable. So this criterion alone is not enough to determine the value of a law. Confidence is a criterion that will have the value between 0 and 1. If the criteria for a rule show certainty of 0.98 means that in 98% of cases if left side of rule is true, the right side of rule will be true too.

$$(A \rightarrow B) = \frac{SUP(A \cup B)}{SUP(A)} \text{Confidence}$$

2.4 Evaluation Criteria of Clustering Algorithms

Evaluation of clustering algorithms is divided into two categories. A set of indicators that are internal or unsupervised that determine clustering operations quality with respect to the contained information in the data set. Other categories that call foreign or observers, according to information out of the analyzed dataset, evaluated the performance of clustering algorithm. In this study, a criterion of unsupervised is used. The criteria Average Silhouette Coefficient that is abbreviated ASC.

As we know the duty of a clustering algorithm is to minimize the distance of inside the cluster with density (coh) and to maximize the distance between cluster with separation (Sep). Because there are many unsupervised criteria, the criteria define two factors in a certain way. ASC criteria define these two factors as follows:

$$Coh = \frac{1}{m_i}\sum_{y\in ci} dist(x,y)$$

$$Sep = \min_{j\le nC_{j\neq 1}} \left\{ \frac{1}{m_i}\sum_{y\in ci} dist(x,y) \right.$$

So, ASC or Silhoutte Measure defines as follows :

$$\mathrm{ASC} = \frac{1}{\mathrm{n}_c}\sum_{i=1}^{n_c} \frac{Sep(i)-Coh(i)}{\max(Sep(i)-Coh(i))}$$

The maximum amount for this measure is the number 1 and the minimum is -1. In the above formula dist (x, y) represents the distance of the (record) x, y of each other that to calculate it, the Euclidean distance is used. Also nc, mi, ci respectively represent the i-th cluster center, the number of cluster member i and the total number of formed clusters for the study points. Euclidean distance is as follows:

$$d_e(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

In the above equation, n represents the number of features (of the problem), y_k and x_k respectively the kth are features of two records x and y.

2.5 Method

In this study, two algorithms rule base Apriori, Fp Growth and three clustering algorithm are used K-Means, two-step (Two Step) and Kohonen. The results are compared to each other and the best rules and the most suitable characteristics is announced and extracted from each algorithm after consultation with specialists and experts of insurance that specify the loss of a cluster.

3. The Rule-based Algorithm (The Discovery of Association Rules)

Implementation of these algorithms according to the methods of working with the used software like all other algorithms, first data source enter to software and then have to run the algorithm. The algorithm was run with the default software parameters and different parameters were applied to the algorithms that the best response was to run with the default software settings.

3.1 FP Growth Algorithm

The rules of this algorithm are as shown in Table 6.

Table 6. Extracted rules by the algorithm Fp growth

Rule	Result	support	Confidence
Used = trolleys, Systems = Nissan discount no damageless than 1.5 million rials, the type of vehicle = rides, year of manufacture more than 2007	Damage	6%	38%
	Damage	47%	40%

3.2. Weka Apriori Algorithm

The rules of this algorithm are as shown in Table 7.

Table 7. Extracted rules by Weka Apriori algorithm

Iteration	k	%train Partition	% Test Partition	Partitioning
8	9	10	90	YES

3.3 Clustering Algorithm

The objective of this part is use of clustering algorithm K-Means, Kohonen and two-step data and check on whether these algorithms on the data will have good output or not? After running the algorithm, the output will be evaluated with criteria ASC.

Implementation of the algorithm according to the method with used software applied various parameters for algorithms that the best response was to run with the parameters that are described below.

3.4 K-Means Algorithm

The best obtained performance for this algorithm has been by setting the following parameters (Table 8):

Table 8. K-Means algorithm parameters settings

%train Partition	% Test Partition	Partitioning
40	60	YES

As shown in Figure 1, after the 8-order of algorithm implementation achieved to zero error percent.

Number of clusters: 9

Iteration	Error
1	1.96
2	0.553
3	0.429
4	0.18
5	0.134
6	0.111
7	0.005
8	0.0

Figure 1. Achieve of the error percent to zero after 8 orders.

Implementation for 9 clusters in K-Means algorithm.

12 more effective fields according to detection of this algorithm for clustering as has been determined in Figure 2.

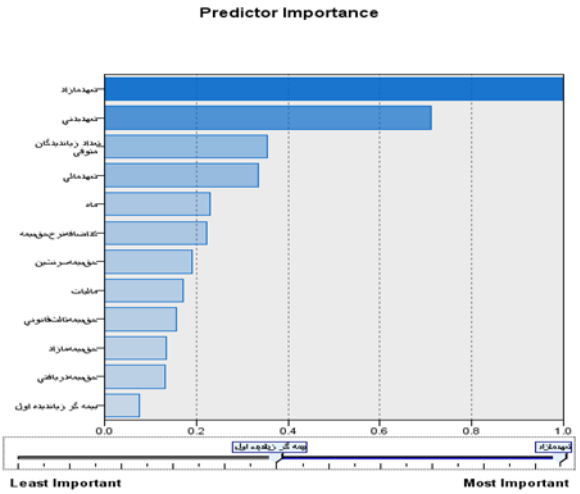
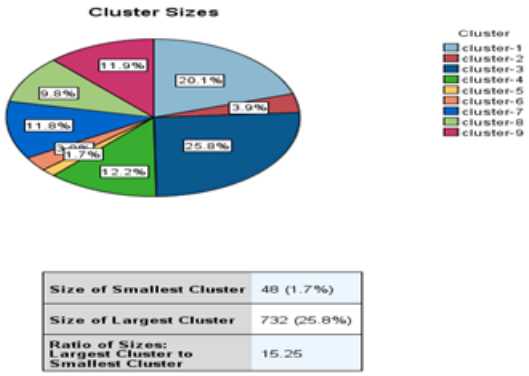


Figure 2. Predictor importance for K-Means.

The fields in order of importance according to the detection of algorithm are (Figure 2):

- Surplus commitment.
- Physical commitment.
- Number of decrease stricken.
- Financial commitment.
- Month.
- Add Code of premium rate.
- Seat premium.
- Taxation.
- Legal party premium.
- Surplus premium.
- Premium.
- First injured insurer.

The size of these clusters is shown in Figure 3.



Size of Smallest Cluster	48 (1.7%)
Size of Largest Cluster	732 (25.8%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	15.25

Figure 3. The size of the clusters and the smallest proportion of cluster to the largest cluster in K-Means algorithm.

Clustering quality has been determined also relatively good as shown in Figure 4.

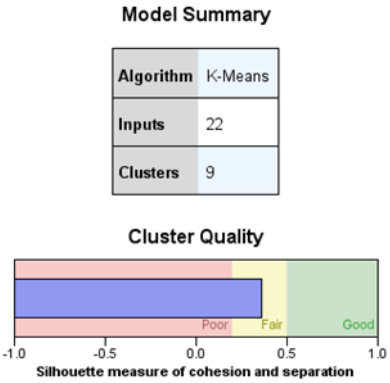


Figure 4. The quality of clusters in K-Means algorithm.

As it is specified, the best determined quality according to criteria of Silhouette Measure has been equal to 4.0, which is also acceptable.

3.5 Kohonen Algorithm

The best obtained performance for the algorithm with parameter settings has been according to Table 9.

Table 9. Kohonen algorithm parameters settings

%train Partition	% Test Partition	Partitioning
40	60	YES

The best number of clusters according to detection of algorithm has been 8 clusters. 12 more effective fields according to detection of algorithm for clustering has been determined as Figure 5.

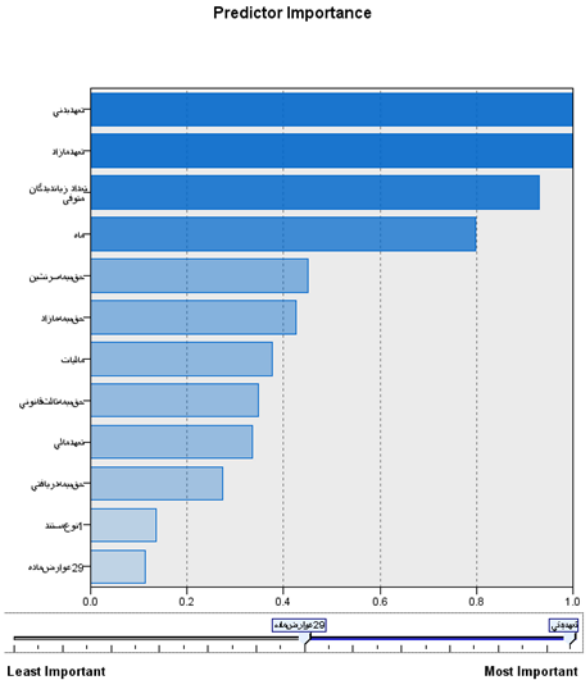


Figure 5. Predictor importance for Kohonen.

The fields in order of importance according to the detection of algorithm are:

- Physical commitment.
- Surplus commitment.
- Number of decease stricken.
- Month.
- Seat premium.
- Surplus premium.
- Taxation.
- Legal party premium.
- Financial commitment.
- Premium.
- TYPE of document 1.
- 29 Article complications.

The size of the clusters is shown in Figure 6.

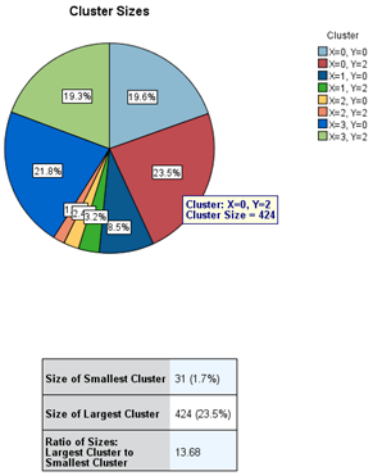


Figure 6. The size of clusters and the smallest clusters to.

The largest cluster in Kohonen algorithm clustering quality also according to Figure 7 has been determined relatively good.

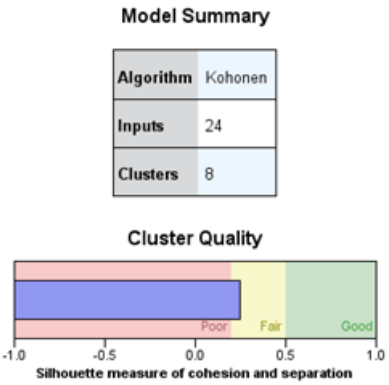


Figure 7. The quality of the clusters in K-Means algorithm.

As it is clear, the best determined quality according to criteria of Silhouette Measure has been equal to 3.0 that is also acceptable. This algorithm is neural network type and therefore the input layer is detected 76 neurons and output layer is detected 12 neurons. (Figure 8).

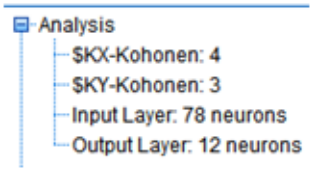


Figure 8. The number of input and output neurons in Kohonen.

3.6 Two-Step Algorithm

The best obtained performance for the algorithm has been with parameters settings of Table 10.

Table 10. A two-step algorithm parameters settings

Algorithm name	Silhouette Measure	The number of clusters
K-Means	4.0	9
Kohonen	3.0	8
Two steps	2.0	3

The best number of clusters according to the algorithm detection has been 3 clusters. 12 more effective fields according to algorithm detection for clustering have been shown in Figure 9.

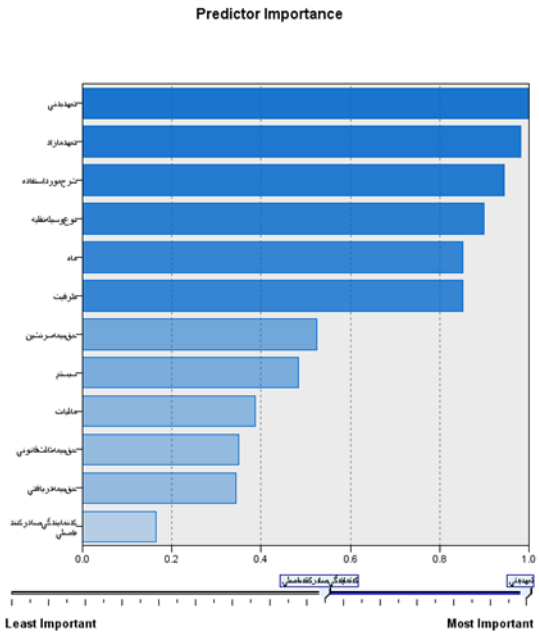


Figure 9. Predictor importance for two step algorithm.

The fields in order of importance according to the detection of algorithm are:

- Physical commitment.
- Surplus commitment.
- More used.
- Vehicle type.
- Month.
- Capacity.
- Seat premium.
- System.
- Taxation.
- Legal party premium.
- Premium.
- Agency code of major exporter.

The size of the clusters is shown in Figure 10.

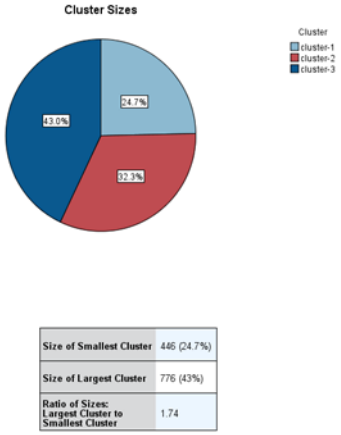


Figure 10. The size of clusters and the smallest clusters to the largest cluster in a two-step algorithm.

Clustering quality has been determined poor as shown in Figure 11.

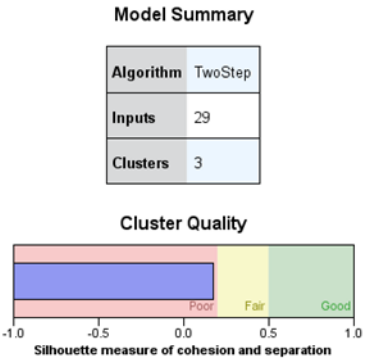


Figure 11. Quality of clusters in the two-step algorithm.

As it is specified, the best determined quality according to criteria of Silhouette Measure has been equal to 2.0, which is less than normal.

4. Conclusions

4.1. Analysis of Clusters Result

In this section we study compare of clustering algorithms.

Table 11. Comparison of clustering algorithms

Late penalty	Physical commitment	Built	Year
Add code of premiumrate	Surplus commitment	Type ofvehicle	System
premium	The number of injured stricken	Term of insurance	Financial commitment
Taxation	Number of cylinders	Employee	Insurance policy of the previous year
Seat premium	First injured insurer	Issued by branch	Plaque
Legal party premium	Capacity	The amount of damage	More used

The results show that the algorithm of K-Means has formed the best clustering with 9 clusters that have relatively good quality. It means that has been able to maximize the distance between the cluster and minimize the within cluster distance. Then Kohnen could generate 8 clusters with low quality that is acceptable also. And at the end, the two-step approach has produced 3 clusters with poor quality. What has concluded that existence of 8 or 9 clusters has been the best number of clusters for this type of data.

In K-Means algorithm, first, eighth cluster with coverage of 10% and then second cluster with coverage of 4% of total cluster have the most similarity with the damage cases, because the add code had the rate more than 30, have low premium and paid a high late penalty.

In Kohonen algorithm, cluster of X = 0, Y = 0 has the most similarity to the damage cases because the add code had more rate than 30, have the low premium and paid a high late penalty. This cluster is allocated 20% of the clusters. Two step algorithm in cluster 1 with 25% of the total volume had the most similarity to damage cases because the add code had the rate more than 30, the auto was Peikan and paid a high late penalty.

So any new records after comparing to these algorithms can lead to damage in future if granted to clusters by the said possibilities that might lead in the future to cause damage.

Clustering algorithms have identified 12 effective features in clustering. These features are marked as underlined in the following table.

According to the results we can conclude that clustering algorithms able to create a new model and approach to the allocation of new samples to a specific cluster to determine the possibility of lose a policy.

4.1 The Result of Correlation Rules (Rule Base)

The best obtained results of the algorithms that obtained rules of A and then B for 3.154 records are according to the following Table 13.

Due to the low of support obtained rules in accordance with the stipulated scientific criteria in the above table, the rules are not reliable scientific and conclusions.

Table 12. Obtained fields of clustering algorithms

Line	Rule	Result	Support	The number of records, including betting and results	The number of records including betting	Confidence
1	No surplus commitments, Add Code of premium rate = 0	Damage	32%	312	755	41%
2	Used = trolleys, System = Nissan	Damage	6%	51	135	38%
3	Discount no damage Less than 1.5 million rials, vehicle type = Peikan, Year of built more than 2007	Damage	47%	437	1090	40%

5. Recommendations

In this study the defects and shortcomings of the current procedure entry of insured and the injured information determined to a certain extent. The approach to the losing policy holders and no harm that is done now had some defects that by fix it, gave them more profit the insurance companies. Accordingly, it is proposed:

- Insert the insured individual characteristics such as age, occupation, education, date of certification issued, type of certification or an individual health position in insurance policy issued for future use of data mining that definitely will lead to find more definitive knowledge in this field.
- Insert more detailed information about the accident, the scene and damaged personal information and responsible for future use of data mining.

6. References

1. Long L, Liang C, Hui Y. Efficient evolutionary data mining algorithms applied to the insurance fraud prediction. International Journal of Machine Learning and Computing. 2012 Jan; 2(3):308–14.
2. Patil SP, Patil UM, Borse S. The novel approach for improving Apriori algorithm for mining association rule. World Journal of Science and Technology. 2012; 2(3):75–8.
3. Ramamohan Y, Vasantharao K, Chakravarti CK, Ratnam ASK. A study of data mining tools in knowledge discovery process. International Journal of Soft Computing and Engineering (IJSCE). 2012 Jul; 2(3):191–4.
4. Saniee M. Applied data mining. First Printing. Tehran, Iran: Niyaz Danesh Publishing; 2012.
5. Allahyari R, Vahidy K. Applying data mining to insurance customer churn management. 3rd International Conference on Information Computing and Applications (ICICA 2012); Chengde, China. 2012 Sep. p.14–6.
6. Firuzi M, Shakoori M, Kazemi L, Zahedi S. The identification of fraud in auto insurance using data mining method. Journal of Insurance (Insurance Industry of the Former). The Twenty-Sixth Year. 2011 Sep; 3(103):103–28.
7. Heydari N, Samrand K, Farahi A. The classification of the insured risk of auto insurance using data mining algorithms. Journal of Insurance (Former Insurance Industry). The Twenty-Sixth Year. 2011 Jan; 104:107–29.
8. Montazeri-Gh M, Mahmoodi-k M. Development a new power management strategy for power split hybrid electric vehicles. Transportation Research Part D: Transport and Environment. 2015 Jun; 37:79–96.
9. Delafrooz N, Farzanfar E. Determining the customer lifetime value based on the benefit clustering in the insurance industry. Indian Journal of Science and Technology. 2016 Jan; 8(1):1342–9.
10. Montazeri-Gh M, Mahmoodi-k M. An optimal energy management development for various configuration of plug-in and hybrid electric vehicle. Journal of Central South University. 2015 May; 22 (5):1737–47.
11. Jeon Y, Lee J, Kwon, D. Process innovation case study of insurance industry: Based on Case of H Company, Indian Journal of Science and Technology, 2015 Jan; 8(S1):20–7.