Available online at www.elixirpublishers.com (Elixir International Journal)

Information Technology

Elixir Inform. Tech. 88 (2015) 36397-36405

ISSN: 2229-712X

Examining the Possibility of Rearrangement of Job Classes Based on Risk in Individual Accident Insurance using Data Mining Techniques

Mohammad Hassan zadeh[1] and Neda Ghazimoradi[2]
[1]Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran.
[2]Master's degree in Information Technology Management, Tehran, Iran.

ARTICLE INFO

Article history:
Received: 16 September 2015;
Received in revised form:
10 November 2015;
Accepted: 16 November 2015;

Keywords
Risk,
Individual Accident Insurance,
Data Mining,
Neural Networks.

ABSTRACT

Incorrect classification of risks or Policyholders can lead to the wrong choice for insurer, and in individual accident insurance, analyzing risk and determining premiums are based on the individual jobs in a job class. Therefore, the risk of the false classification of risks and Policyholders can lead to significant financial losses for insurance companies or policyholders. So in order to recognize and optimize the classification of job`s risk and thus determine a reasonable price for the policy insurance, the frequency and severity of losses can be considered as determining variables. In this study the collected data from Dana Insurance Company database, were entered in the SOM neural network in the form of three variables to assess the occupational risk. After reaching the overview of clustering resulted from SOM, and determining the optimal number of clusters based on Silhouette index, the separation of final clusters using different algorithms, K-Means and TwoStep, was done. Finally, the precise rules found for describing the outputs of the algorithms based on C5.0 algorithm with accuracy of 100%. Due to the overlapping clusters, there was no possibility of allocating jobs to different occupational classes. There is also another notable result and that is the significant negative correlation between the two factors, the current risk level and the damage ratio (the ratio of claims paid to premiums received).

Introduction

Occupational Classification of Individual Accident Insurance is the most important issues to be discussed. This is because the basis for analyzing risk and determining Premium are Jobs in this category.

One of the major issues in the insurance industry is to understand the factors and variables that are important for predicting the probability and amount of damages. One of this effective factors on risk which is the most important element in issuing Individual Accident Insurance is job(Main job and secondary jobs) in order to determining the premium rates and also assessing the probability and amount of damages [1]. According to the available loss data and the opinions of experts seem that available jobs in various classes need to be reviewed.

Jobs are divided into 5 categories in Individual accident insurance that each of these classes represents the position of individuals in terms of the kind of risks that a person is faced with them in the workplace. (The first class includes the least dangerous jobs and the fifth Class is the class of the most dangerous jobs). In other words, the main criterion in the appropriate selecting of risk and therefore risk management is job classification. If the insurance industry fails to identify and manage risks properly, will be deprived of the possibility of surviving and continuing the operations. But there is the possibility of assessing the frequency and severity of damages resulted from the available occupations in these classes based on the available loss data. So the main strategy of insurance companies is implementing an effective mechanism facing with the diversity of risks which annually take place in these companies [1]. With increasing knowledge and awareness of the risks - which in this study is referred to as jobs classification - it will be easier for insurance companies to calculate their costs and revenues. And also they can set their annual budget based on real costs. Additionally, determining the appropriate, reasonable and fair insurance rates would be possible. Accurate estimate of the expected loss and frequency and severity distributions, are the basic operation for efficient decision making in the risk management [2].

The occupational classification of Individual Accident Insurance needs to review and re-review since from the beginning of the plan and its implementation many years have passed, and unfortunately, no comprehensive study has been done on this issue. While social, political, economic and technical changes usually affect the frequency and severity of risks. Hence, there is a continuing need for analyzing the information and assessing the impact of the changes.

This research tries to use data mining techniques such as classification, clustering and neural networks using the data of issued and damaged insurance policies contained in the database of Dana Insurance Company, in order to identify and classify the jobs in Individual Accident Insurance again. Also by determining the real price of this insurance policy moves towards greater profitability and determining the new strategies.

In the following, in section 2, the theoretical foundations, section 3, research methodology, section 4 the research findings and finally in Section 5 the conclusion and the recommendations will be discussed**.**

Theoretical Foundations

"The Insurance is one of the main risk management tools that play an important role in the economic, social and political life in countries." [3] Although a lot of definitions have been proposed for risk, in 1966, the commission on Insurance

Tele:
E-mail addresses: nedaghazimoradi81@yahoo.com

Terminology of the American Risk and Insurance Association approved the following definition of risk: “uncertainty as to the outcome of an event when two or more possibilities exist.” (American Risk and Insurance, 1966).From this definition, one can easily conclude that the greater the uncertainty, the greater the risk. Regardless of how risk is defined, the effects of risk affect the economic performance of factors. Then it imposes limits to the optimal allocation of resources and the economic development of countries. Individual decisions as business decisions are not made under conditions of certainty. Although the visualization of risk idea may be difficult, all economic agents make decisions that think they are profitable for them [4]. In this section, the theoretical foundation related to risk management processes, individual accident Insurance as well as data mining techniques such as self-organizing networks, K-Means and TwoStep algorithms, and also C5.0 algorithm are studied.

The process of risk management

The risk management process consists of four main steps are as follows:

Phase I - Identification of risk

Perhaps one of the most important tasks of the risk management is the process of identifying. The inability to identify one of the potential several events may lead to financial bankruptcy. It is important to anticipate problems before they materialize. But there is no scientific method or systematic approach to the process of identifying. The risk identification is defined as "the systematic, continuous process that identifies the properties, liability and persons at risk upon the occurrence or earlier." [3]. Analyzing of job or position is other underlying to identify risks.

Phase II - Measuring of risk

Once risks were identified a risk manager will measure them. Risk measurement is the process of creating data. This means assessing the potential size of losses and how much the probability of it is. The required data are related to the two dimensions of risk: the frequency and severity of events that will occur. The Information can be collected in order to describe, assess, and prevent a phenomenon.

Phase III - Risk assessment

Data related to the frequency and severity of injury, do something beyond the recognition of important damages. These data are also very useful to determine the way of dealing with the situation. Purposes of risk assessment are to determine the seriousness of the risk and also determine whether people have been exposed or community.

Phase IV - Risk control

whether the firm has decided to accept the risk or not, it should consider two possibilities: reducing the frequency of events as well as the severity of damages that will occur. The potential benefit resulting from risk control should be weighed against the costs. If the expected benefit is not equal to or greater than the costs of those activities, it is better that the firms not to venture [5].

Individual Accident Insurance

In this type of insurance persons covered by individuals accident insurance against occupational and non-occupational or personal threats at all times, 24 hours a day for a year [6].

Accident

The accident is a sudden and severe event caused directly by an external force and without the will of the insured resulting in bodily injury or damage [6].

In individual accident insurance, the factors for determining rates are insured job and activities outside of his job. Because of the large number of jobs, they are divided into distinct classes that this classification is according to the severity of their occupational risks as described in Table 1.

Table 1. Classification of occupations based on the occupations risk [1]

Example	Risk	Job class
Such as office workers and people who work in the office just to provide the service.	People who are in daily activities with minimal risk.	**First class**
Doctors, dentists, engineers, supervisors, marketing, knitters and warehouse keepers	People in their daily activities compared with subjects in class one, have been faced with more relative risks and In addition to the use of intellectual power, usually work with their hands but will not work with industrial machines.	**Second class**
Farmers, drivers and construction workers	Includes persons who are skilled or semi-skilled, and most of them usually work with machines and industrial equipment.	**Third class**
Pressing milling, mast section, fireman, sweeper and Dock	People who work with high-risk industrial machinery and equipment or the type of work they are doing is dangerous. Unskilled workers in industries that face with many tasks are in this class.	**Fourth class**
Pilots and miners	People in their daily activities faced with the greatest risk.	**Fifth class**

Types of risks covered by the insurance

4 types of risk covered by this insurance type are:

Insured's death

If the insured dies due to the risks covered by the insurance, insurer is committed to pay the anticipated lump sum capital of the insurance policy.

Disability and permanent total disability

In case of occurring an accident or the total permanent disability, the insurer will pay all the committed capital to the insured.

Permanent disability

If the insured suffers disability due to an accident, according to the table of disabilities and impairment, disability compensation will be paid to the Insured.

Compensation for medical costs

If the insured injured in an accident and incurred medical expenses, the insurer would compensate the medical costs up to the committed amount [7].

Risks covered by insurance: Risks that are normally covered by the accident insurance policy and if it happens, the insurer will undertake to pay compensation or medical expenses. These risks are: the risks of vehicles accidents, rescuing persons in danger, fires, explosion, collapse, dangers of lightning, poison of gases and vapors, rabies disease, tetanus, anthrax and etc [7].

If the insured do things like hunting, riding, sailing, scuba diving and skiing and etc that have more risks addition to their professional normal activities intermittently or continuously, regardless of the insured occupation, it could be possible to cover each of the activity with specified additional rates. If a person wants to be insured for one or more specific risks, extra premiums that has to pay is not the sum of extra premiums but the highest additional premium received from him [1].

In general, the insurer believes that misclassification of risk or misclassification of policyholders leads to incorrect selection. Insurance companies are aware of the possibility of bad choices, so try to protect themselves by adding the terms and conditions of the insurance policy [3]. Since the basis of risk analysis and determination of premiums in this insurance is the individual jobs in the job classes, so the misclassification of risk and insured can lead to considerable financial losses for insurance companies or even policyholder. Because the incorrect assessment of each job and allocate it to the inappropriate class, the determined premium rate would not be appropriate, reasonable and fair. Thus, the frequency and severity of the damage can be detected as determining variables in the optimal identification and classification of job`s risk as well as determining the reasonable and realistic price of the insurance policy.

Research Methodology

Self-organizing Map networks (SOM)

Artificial Neural Network (ANN) is defined as an information processing system that has characteristics resembling human neural tissue. The existence of ANN provides a new technology to help solve problems that require thinking of experts and computer based routine. A few of ANN application was for classification system (clustering), association, pattern recognition, forecasting and analysis[8]. And in this paper, the ANN method which applied is SOM method, that to be used to achieve an overview of basic clustering of occupational hazard classes.

SOM is an unsupervised neural network that is formed by neurons in a regular neural network structure in low dimension. Each neuron has an n-dimensional weight vector that n, is equal to the dimension of the input vectors. Vectors of weights (synapses) connect the input layer to the output layer (it is called competition layer). Neurons are interconnected by a neighborhood function. Each input vector actives a neuron in the output layer called the winner cell, based on the maximum similarity. The similarity is usually measured based on the Euclidean distance between two vectors. The main difference between SOM training algorithm with other vector quantization algorithms is that In addition to the communication weight of the unit with the highest adaptation (winner neuron), the weights of the neighboring cells of the winner cell are updated. Close Observations in the input space, actives the two close units in the map. Training phase continues until the weight vectors are stable and do not change anymore.

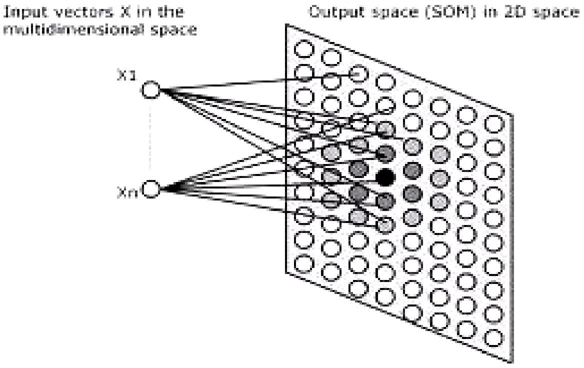
Input vectors X in the multidimensional space
Output space (SOM) in 2D space
X1
Xn

Figure 1. The structure of self-organizing map neural network [9]

After the training phase, this neural network will be able to cluster n samples in the form of m unit of the map (the cluster). SOM now must be given as input to K-means algorithm [9].

Clustering

Cluster analysis is a method for grouping data or observations with respect to their similarity or proximity that with it the data or observations are divided into distinct homogeneous groups. Clustering is one of the unsupervised learning methods. Classification and clustering are different. In classification, each data set is allocated to a class or classes, but in clustering, there is no information on existing classes within the data and clusters are extracted from the data [10]. Cluster analysis procedures include: Selecting criteria similar or close observation, Selection of cluster analysis, the decision about the number of clusters, Interpretation of the categories or groups formed.

Clustering algorithms

Clustering data algorithms are examined in terms of three general approaches: Hierarchical clustering algorithms, Partitional clustering algorithms, Fuzzy clustering algorithms

Hierarchical clustering algorithms

In the hierarchical clustering methods, the hierarchical structure of the data is attributed to the final clusters based on their popularity. The hierarchical structure of hierarchical clustering methods is said the dendogram graph. Types of hierarchical clustering algorithms are: Average-link, Group average link, Complete-link. The main difference between these methods is how to calculate the similarity between clusters [10].

The Partition Clustering

What is obtained from a Partition clustering algorithm is a single group of data instead of a cluster structure. The process of producing clusters in Partition methods is that an assessment function already defined optimizes the process and in this case calculating the optimal value for it is expensive.

Thus, in practice, these algorithms are usually run several times on the sample and the best combination is used for the output of the clustering [10].

The best known partition algorithm is “K-means “which is the basis of many other clustering algorithms [9].

The Fuzzy Clustering

In traditional clustering strategies, each sample belongs to one and only one cluster. Thus, the clusters are well separated and detached. In the fuzzy clustering methods, each sample can be attributed with a membership function to each of the clusters. A member can also be a simultaneous membership in two or more clusters considering the membership function [10]. In this study the K-MEANS and TWOSTEP algorithms are used as a kind of partitioning and hierarchical algorithms.

The clusters evaluation criteria

There are many clustering evaluation criteria that the most important of them are: Dunn index, the index Davies-Bouldin index and Silhouette Coefficient that the Silhouette Coefficient is used in this study

The Silhouette Coefficient

It measures the ratio of the average distance of each observation from its own cluster or category to the nearest cluster or other categories. The Silhouette Coefficient is calculated according to the following equation:

$$s_i = \frac{a_i - b_i}{\max\,(a_i\,, b_i)} \qquad (1)$$

In this equation b_i is the average distance of (i) from the cluster to which it belongs, and a_i is the average distance from the nearest cluster apart from the own cluster. This measure simultaneously evaluates the coherence within clusters as well as the separation between the clusters. Silhouette index represents the average between 1 and – 1 based on the amount of variance between and within clusters and the closer it is to 1, it means the optimal point of the index. At this point, the dispersion between the clusters is maximized and the distribution within clusters reaches its minimum [11].

Clustering with the K-means algorithm

Algorithm K-means divides the data set into k subsets (clusters) so that all components of each subset have the minimum distance to the cluster center. By this algorithm k input samples are randomly selected as cluster centers. Then it assigns the rest of the input samples, based on the minimum Euclidean distance with the determined cluster centers to the appropriate clusters. After that the mean of each cluster is recalculated and considered as a new center of cluster. This operation is repeated until the cluster centers do not change anymore. The criterion that must be minimized in K-means is:

$$E_{K-means} = \frac{1}{C}\sum_{k=1}^{c}\sum_{x\in Q_k}\|x - C_k\|^2 \tag{2}$$

In the above equation, C is the cluster numbers, Q_k is the K th cluster and Ck is the center of the cluster Q_k. [9]

This algorithm requires the user to determine the number of clusters. Therefore, this method requires a lot of trial and error. As a result, the analyst must use the other automatic clustering formulas to obtain an overview of the number of clusters [12].

The two-stage clustering method (TwoStep)

This method is useful when you want to classify the data in different clusters, but have no idea of the number of categories. Like other clustering methods, this method does not select any data as the purpose but divides the data into clusters with the highest similarity within clusters and low similarity between the data of each cluster with the data from other clusters [13]. This kind of clustering is a two-step procedure. In the first step, an analysis is performed on the data and the data is divided into manageable sub-clusters. In the second step it clusters the data through the method of hierarchical clustering with regard the first step and makes larger clusters [13]. One of the advantages TwoStep algorithm is that it provides facilities for handling outliers [12]. In Table 2, an overview of our 3 selected algorithms in this paper is presented.

Classification

The classification is the process of finding models that by distinguishing classes or concepts of data can predict the other unknown objects.) Classification is a learning function that maps the data to one of the predefined categories. The data are divided into two parts: training and testing. Training data are used by the system to learn the rules. Test data are used to verify the accuracy of classification and prevent over-fitting [14].

Some common methods of classification are: The decision tree (the decision tree algorithms such as: C5.0 C4.5, CART), Bayesian classification (simple Bayes and Bayesian networks), back propagation neural network, Support Vector Machines, Association classification, sloth Learners (nearest neighbors, Case based reasoning) and other methods such as genetic algorithms, imprecise sets, and fuzzy sets [15].

C5.0 Classification algorithm

This algorithm is a univariate decision tree algorithm that is improved C4.5 algorithm which was designed in **1993** by Australian researcher J.ross Quinlan. These algorithms such as CART algorithm initially provides almost a full tree, but its pruning strategies are quite different. C5.0 algorithm divides the data into sub-collections that contain the records which are more homogeneous than their parents. In C5.0, dividing the samples is done based on the field that has the most information. This algorithm selects a branch of the fields for branching with the reducing rate of unknown data based on the following equation.

$$H(X \mid a) = -\sum_{j}\sum_{i} p(a = a_j)p(C_i \mid a = a_j)\log p(C_i \mid a = a_j) \tag{3}$$

(a) is the value of a field, (C) is the class label and H (x.a) is a field with a minimum value of H (x.a) is selected as a branch. Each sub-sample is determined by the first branch. Then again, the segmentation is done usually based on another field and this process is repeated many times until the samples are not dividing. Finally, the lower branches are re-examined and worthless branches are eliminated [15].

As shown in Figure 2, in this paper, after reviewing the literature, considering the severity and frequency of risk assessment criteria, and using data mining algorithms and neural networks in accordance with research subject, and also through modeling, we assess the levels of risk and, if possible, provide the optimal classification.

After reaching overview of the clusters resulting from the application of SOM neural networks, and determining the optimal number of clusters based on the silhouette index, the separation of the clusters is done by comparing the results of using the hierarchical and partitioning algorithms (K-Means and TwoStep). Achieving precise rules for describing the output of the algorithm will be done by setting the training and test data (in both K-Means and TwoStep algorithms) and by using the C5.0 decision tree algorithm.

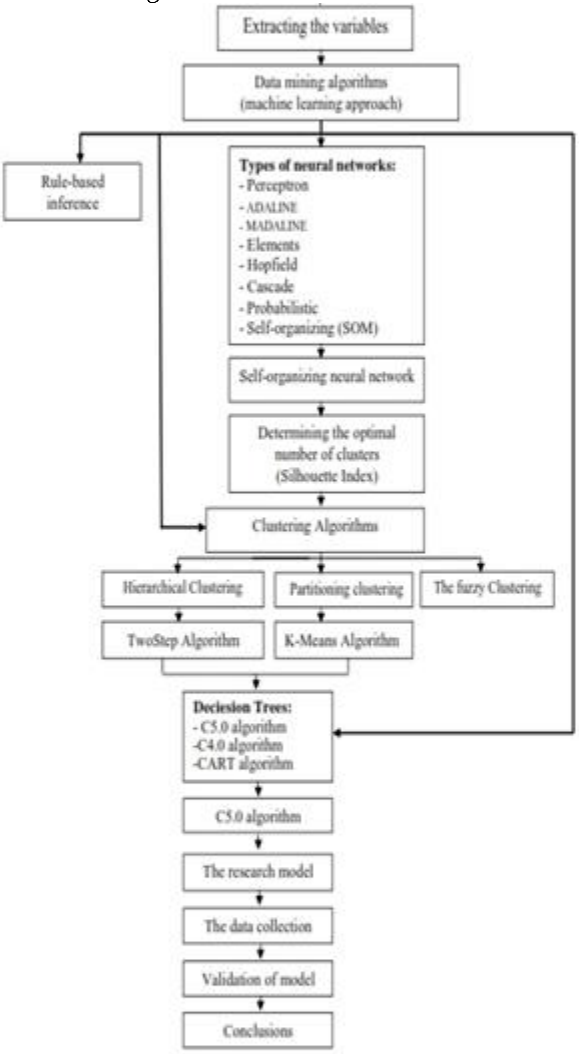

Figure 2. Conceptual framework of the study

Research Findings

The statistical population used in this study includes data on insurance policy issued for the 59,562 insured and the 1052 damaged insured that Which in the interval 3/21/2007 until 9/20/2012 have been received from the database of Dana Insurance company. Due to the nature of the data mining that with more data, gives better results and In order to obtain results with higher reliability all data in the database are used and sampling has not been done.

In Dana Insurance Company's database, the total number of jobs registered for determining the job classes are 4191. 670 of these jobs, about the 59,562 insured that have bought this type of insurance during 2007 till 2012. But only the owners of 132 existing jobs have been damaged in this time. Part of the data in the database contains the following fields: Insurance number, gender, date of birth, occupation, occupational hazard classes, the amount of the total premium, the type of transfer, the amount of payable compensation and the percentage of the minor disability. To achieve this goal, we need to the data cleaning, data Reduction and the data transformation so many of the 132 jobs combined and integrated due to the same nature and only slight differences in writing the names. So that eventually 84 jobs remained. Because of using the Clementine software, we do not need to change the scale to transform the data (standardizing, changing the scale of data and feature construction). Therefore only qualitative variables of type of claims (death - Impairment - medical costs) and gender became quantitative.

Network inputs

Type of claims (death - Impairment - medical costs) percentage of disability and job loss ratio that measure the frequency and severity of the jobs risk, are considered as inputs to the network. It should be noted that the variable of claim types became quantitative based on a spatial bipolar scale and insurance experts opinions, by assigning the number 9 to death, 7 to impairment and 3 to medical costs.

Implementation steps

Initially, the input variables enter the SOM neural network for the initial clustering. Determining the learning rate is one of the influencing factors in the training process of the network. The type of this reducing learning rate can be selected the linear or exponential.

Thus, initially, the total data enter the network in the form of the 3 variables and with selecting 3 neurons in the input layer and 12 neurons in the output layer (default setting), 6 initial clusters were generated by the network. Since there are some occupations that have few records and was led to the formation of sparse clusters, and also in this case most of the clusters are dependent to the two variables of type of claims and the impairment percentage and the standard deviation of the cluster is also very variable and high, so records of jobs with fewer than 10 were eliminated. So that in the end, 22 jobs of 84 - which included 822 people injured- were left.

It should be noted that 22 of these jobs, covers 48,065 of the 59,195 insured person who bought the individual accident insurance. It is equivalent to 81% of the data that statistically, the results will generalize to all the data.

Again, the 822 remained records enter the network in the form of the 3 variables and in three iterations with regard to the states of linear and exponential learning rate, and default settings.

Table 2. Comparative table of clustering techniques [12]

Algorithms	**K-Means**	**TwoStep**	**Kohonen network/SOM**
Methodology description	Iterative procedure Based on a selected (typically Euclidean) distance measure	In the first phase records, through a single data pass, are grouped into pre-clusters. In the second phase pre-clusters are further grouped into the final clusters through hierarchical clustering	Based on neural networks. Clusters are spatially arranged in a grid map with distances indicating their similarities
Handling of categorical clustering fields	Yes, through a recoding into indicator field	yes	Yes, through a recoding into indicator fields
Number of clusters	Analysts specify in advance the number of clusters to fit; therefore it requires multiple runs and tests	The number of clusters is automatically determined according to specific criteria	Analysts specify the (maximum) number of output neurons. These neurons indicate the probable clusters
Speed	Fastest	fast	Not so fast
Integrated outlier handling (in IBM SPSS Modeler	No	Yes	No
Recommended data preparation	Standardization and outlier handling	Standardization and outlier handling	Standardization and outlier handling

The size of the network map and parameters like the radius of the neighborhood and the number of cycles are presented in table 3.

Table 3. learning models and the cluster number

Learning model	**Dimensions of matrix**	**First stage of training**	**Second stage of training**	**Number of clusters**
Linear learning rate	**10×7**	Neighborhood :4 Initial Etha:./4 **50** Cycles:	Neighborhood :2 Initial Etha: ./25 Cycles:250	**26**
Exponential learning rate	**10×7**	Neighborhood :4 Initial Etha:./4 Cycles:50	Neighborhood :2 Initial Etha:0/25 Cycles:250	**24**
Learning with the default settings		Neighborhood : Initial Etha: Cycles:	Neighborhood : Initial Etha: Cycles:	**7**

After several running the network with changing the different parameters and in the state of the linear learning rate, we find that in this state we have the best result and the standard deviation as well as the distribution of the main clusters show detailed breakdown of the data in main clusters. (SD in main clusters is between 0 to 2 percent). Moreover, in this state, we have 26 clusters that among the 12 clusters that have the bulk of the data, the actual number of main clusters will be four which include 75% of the data.

In the exponential learning rate model, with regard to the parameters of the state, the standard deviation of the variables within the clusters that have the main part of the data is zero. Also in this state, 12 clusters out of 24 accounted for the bulk of the data that of these only the 4 main clusters have the 70% of data. Also with the default settings, only 3 clusters were accounted for 85% of the data. Thus considering the output of SOM neural network based on the different parameters, we can conclude that most of the data have been replaced between 3 to 12 clusters. Therefore, to determine the optimal number of clusters we need an indicator to evaluate and determine the exact number of clusters. To do this, we use the standard silhouette.

cluster number	**Silhouette**
3	0.7556
4	0.8697
5	0.9009
6	0.6722
7	0.6596
8	0.6714
9	0.8053
10	0.8335
11	0.8967
12	0.8512

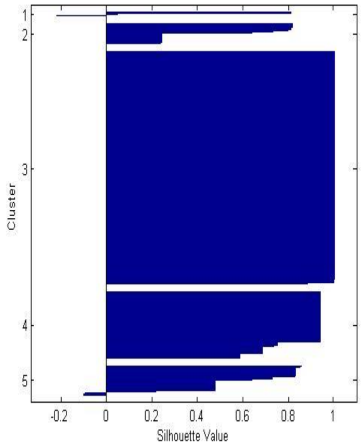

Figure 3. The average silhouette index in different clusters

As seen in figure 3, the average of this indicator has reached its maximum (0.9009) at 5 clusters. As a result in the next stage, using the K-Means algorithm- which gives us the possibility to determine the number of clusters- we do the final clustering.

According to the information contained in Figure 4, various jobs assigned to different classes and the significant point is the repetition of jobs in various classes. With respect to the third feature of the type of claims, percentage of disability and the job loss ratio as the only indicators of assessing the intensity and frequency of risk, there is no possibility to assign each job to one class.

Also to compare and ensure the accuracy of the results we used another algorithm named TwoStep for clustering the data. In this mode and with selecting five clusters, the 3 variables were considered the most important.

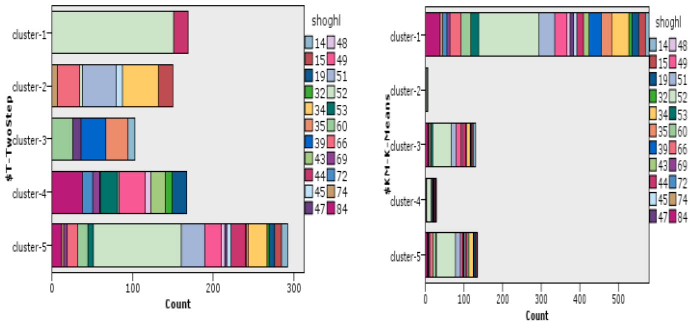

Figure 4. Job`s distribution in clusters resulted from K-Means and TwoStep algorithms

But to describe the clusters and achieve rules for a better understanding of the clusters, we use the C5.0 algorithm and this is achieved by allocating 75% of the data for training and 25% to test.

By applying the C.5 algorithm on the clusters resulted from the K-Means algorithm, there is only the possibility of accessing to the laws that describe the various clusters but it doesn’t represent allocating a particular job to a specified class. Also in this case, the variable of the job losses rate is not considered.

Also by applying the C.5 algorithm on the clusters resulted from TwoStep algorithm, two variables (percentage of disability and job loss ratio) were assessed important. Also, in providing a list of rules, unlike the K-Means algorithm, the variable of the job losses rate has not been considered. Moreover, in this state, there is not the possibility of assigning the specific job to specific clusters too and each job according to different conditions, can be placed in different clusters.

Results also show that the algorithm K-Means states the accuracy of the Train data, 100% and Test data 54/99%. Also, the accuracy of the Train and Test data in TwoStep algorithm is 100%. It means that our patterns for describing the different clusters are very accurate.

Therefore, by investigating the mean and standard deviations for each of the input features to the K-Means and TwoStep algorithms and also by evaluating the rules resulted from applying C5.0decision tree algorithm, the clusters can be matched with the current occupational hazard classes in the individual accident insurance so that the clusters 3, 1, 2, 4 and 5 in the TwoStep algorithm are equivalent in the low-risk to high-risk (1 to 5) clusters. Also the clusters 1, 5, 4, 2, and 3 of the K-Means algorithm, include low risk to high risk clusters.

So it seems that although there is not a significant difference between the two algorithms of K-Means and TwoStep in terms of clustering and describing the clusters by the decision tree algorithm, but giving that when making rules with using the TwoStep algorithm, the three variables are considered and also accuracy is 100% so, the criterion for deciding will be this algorithm. Table 4, shows the rules attributed to the clusters in TwoStep algorithm in compliance with the occupational hazard classes.

Table 4. The Rules Attributed to the Clusters in Twostep Algorithm in Compliance With the Occupational Hazard Classes

clusters	TwoStep Rules	Risk class
1	-Types of claims ≤ **3** -Jobs loss ratio ≤ **0/025** -Impairment percentage≤**2**	**2**
2	-Types of claims ≤ **3** and **0/013**≤Jobs loss ratio≤ **0/021** -Types of claims ≤ **3** and **0/021**≤Jobs loss ratio≤ **0/025** and Impairment percentage> **2**	**3**
3	Types of claims ≤ **3** and Jobs loss ratio ≤ **0/013**	**1**
4	-Types of claims ≤ **3** and Jobs loss ratio>**0/25** - Types of claims>**3** and Jobs loss ratio> **0/052** - Types of claims>**3** and Jobs loss ratio≤ **0/052** and Impairment percentage > **65**	**4**
5	Types of claims>**3** and Impairment percentage≤**65** and Jobs loss ratio≤**0/052**	**5**

The data in Table 5 extracted from Figure 4, shows the percentage of each job belonging to different clusters resulted from the TwoStep algorithm. As can be seen, except for two jobs of plastering building works and carpenter that both of them assigned only to cluster 4, other jobs are distributed in different classes and there is no possibility to allocate a job to a particular class.

Thus it is clearly evident that risk assessment of jobs based on the frequency and severity of risks, does not lead to the breakdown of jobs to certain classes. Therefore, assessing the risk, does not necessarily determine a specific occupational class.

Table 5. The Occupational Distribution in the Clusters Resulting from the Twostep Algorithm

job	Cluster	Belonging each job to clusters (%)	job	Cluster	Belonging each job to clusters (%)
Driver of Light Extremist vehicles	**1**	**%58**	Minibus/bus driver	**1**	**%50**
	5	**%41**		**5**	**% 50**
	2	**%1**	Worker	**2**	**%68**
	4	**%0/38**		**5**	**%32**
manager	**2**	**%70**	Pickup driver	**2**	**%59**
	5	**%30**		**5**	**%41**
Agency driver	**2**	**%62**	Employee	**2**	**%66**
	5	**%38**		**5**	**%34**
Farmer(beneficiary or owner)	**2**	**%69**	Taxi driver	3	**%77**
	5	**%31**		5	**%23**
Student / university student	**3**	**%97**	Housewife	**3**	**%93**
	5	**%3**		**5**	**%7**
Retired	**3**	**%53**	Guard	**4**	**%76**
	5	**%47**		**5**	**%24**
Driver of the heavy Extremist vehicle	**4**	**%76**	Trailers driver	**4**	**%67**
	5	**%24**		**5**	**%33**
Cultivator	**4**	**%67**	Truck driver	**4**	**%61**
	5	**%33**		**5**	**%39**
Welder	**4**	**%73**	- Brick work - stone work -Tile work	**4**	**%72**
	5	**%27**		**5**	**%28**
Seller	**3**	**%65**	carpenter*	**4**	**%100**
	4	**%2/5**	plastering building Work *	**4**	**%100**
	5	**%32/5**			

Also the statistical analysis revealed (Table 6) that the more we move toward jobs in high-risk categories, the amount of loss ratio (ratio of losses paid to received premiums) is reduced. It means that the insurance company will gain more profit from jobs with higher risk. Again, this result indicates the Lack of credibility of the nature of the job classification.

Also there is an inverse correlation between gender and type of claim (intensity). So, men compared to women are more likely to damage severe and more costly claims. Considering that the loss ratio of damaged women to the Insured ones is 8/0 and damaged men to insured men is **1/85**.Age is also another variable that is correlated with the severity of the damage. In this case, the severity of damage increases with aging.

Conclusions and Suggestions

In the Individual Accident Insurance, the main criterion in selecting the risk appropriately and therefore the risk management is jobs classification. While it seems that this classification can`t analyze the acceptable risk. So setting the annual budget of insurance companies based on the actual costs and also determining the suit, reasonable and fair insurance rates would not be possible.

In this study, after reaching an overview of the clusters resulting from the application of neural networks SOM, and determining the optimal number of clusters based on the silhouette index, separating of the final clusters performed using different algorithms, K-Means and TwoStep. So we achieved the precise rules for describing the output of the algorithms with the accuracy of 100% by the C5.0 algorithm.

Table 6. The correlation coefficients between the variables of the job risk class and the loss ratio, gender and type of claim, type of claim and age

		Risk Class	Claim Ratio
Risk Class	Pearson Correlation	1	-.127(**)
	Sig. (2-tailed)		.000
Claim Ratio	Pearson Correlation	-.127(**)	1
	Sig. (2-tailed)	.000	
		gender	death - Impairment - medical costs
gender	Pearson Correlation	1	-.107(**)
	Sig. (2-tailed)		.000
death - Impairment - medical costs	Pearson Correlation	-.107(**)	1
	Sig. (2-tailed)	.000	
		death - Impairment - medical costs	age
death - Impairment - medical costs	Pearson Correlation	1	.142(**)
	Sig. (2-tailed)		.000
age	Pearson Correlation	.142(**)	1
	Sig. (2-tailed)	.000	
	N	1051	1051

Due to the overlap of the cluster, it was not possible to allocate jobs into job classes. So it is clear that Jobs risk assessment based on the frequency and severity of risks does not lead to a breakdown of the jobs to different classes. The risk assessment does not necessarily determine the specific occupational categories.

Also, one of the striking results of this study found a significant negative correlation between the two factors, the current hazard class and the loss ratio (the ratio of paid claims to received premiums). Means with moving towards the high-risk categories, the company is in the profitable status and vice versa for the lower classes, such as third-class that average premium received from the insured and the frequency and severity of the risk is in a bad condition, the company forced to incur large losses on the class. Therefore we do not have enough reasons to get a high premium for the class 5 as a high risk category. This result again shows the lack of the credibility of the job classification nature as well as determining the premium based on them.

Suggestions

[16] by designing a neural network , classified industrial jobs to the high- risk and low- risk classes based on how and the conditions of lifting at work. He also in another study in 2003 investigated the health of employees using various data mining techniques such as neural networks , regression , decision trees and ...and classified the industrial jobs with respect to the risk of work-related back pain disorder. [17] Proposed a classification system for the reason of occupational accidents and by the operational analysis of the events leading to the accidents, their communications and their participation in creating the events, formulated the prevention strategies.

[18] Providing a semi-statistical method of classifying risks, evaluated occupational hazards that the results of these assessments can be implemented to prioritize the work that is being done to control the risk and also classification of occupational hazards in construction projects. Also [19] evaluated the contribution of different factors, such as environmental risks, technical problems, lack of the work organization, how knowledge and occupational knowledge and other human factors in occupational injuries and their relationship with the job, the worker's age and type of events in railway workers and showed that these factors have an important role in the injuries.

In the present investigation, the jobs risk assessment was done based on the frequency and severity of risks using data from the available damage data that did not lead to jobs separation to different risk classes. So, considering the issue that the rules relating to the laws of Central Insurance of Iran, retrieved from regulations over the past half century of European countries that in recent years, many of the factors that influence the realization of the insured risk has changed, such as cases the [19] stated, it is necessary to review and reread the instructions and insurance laws in this field, and try to reduce the occupational classes and remove the two classes. Thus, rather than emphasizing on the fifth classes with boundaries without changing (In the interest of policyholders), the classes can be divided to the three classes of low-risk, medium risk and high risk. In this regard, considering the rules listed in Table 4 as well as the distribution of jobs in different clusters presented in Table 5 can be useful. So that the cluster 3 resulted from the TwoStep algorithm, indicates the low risk occupations, the clusters 1 and 2, lead to the intermediate-risk occupations and the clusters 4 and 5 lead to high risk occupations. As described in Section 2, if an insured in addition to the normal activities of their profession, did tasks such as hunting, riding, sailing, scuba diving and skiing etc. continuously, regardless of the insured job, we can cover them by applying the additional rates.

Also considering the correlation between the two age and sex variables with the claim variable-so that with increasing the insured's age, the severity of damage will increase and there is an inverse relationship between being a woman and the severity of damages- Considering the 2 gender and age variable when issuing an insurance policy, can lead to the appropriate, reasonable and fairly premiums. It should be noted that the lack of the possibility of identifying the cause of the accident was the main limitation of the study So that there was no possibility of separating of the insured job-related accidents than non-occupational accident. As a result, it led to produce the gross qualitative data to some extent.

Appreciating

At the end, I would like to thank all my colleagues in life insurance department to cooperate sincerely and to provide assistance and advice as well as the required data in the present study.

Resources

[1] Mohammad Beigi, A., Amini, M,” Insurance Knowledge”, publication of Iran`s Central Insurance, Tehran, (2012).

[2] Cheraee, J, “Accident insurance and its role in our lives” publication of Iran`s Central Insurance, Iran, Tehran, (2002).

[3] Tirlea M.R., Vlad M.P, “information scenarios over the risk in insurance” Annals of the University of Oradea: Economic Science, Vol 1, Iss 1, (2013), Pp 1044-1050.

[4] Mahdavi, Gh., Nasiri, F, “Theoretical fundamentals of insurance”, Insurance Research Institute Publications, Iran,Tehran, (2014).

[5] Csikósová, A., Mižíková, I., “Insurance as an Important Factor Reducing the Risk in Industry”, Acta Montanistica Slovaca, Vol 14, Iss 3, (2013), Pp 260-267.

[6] Peikarjoo, K., Hosseinzadeh, L.,“Specialized dictionary of insurance terms”, Insurance Research Institute Publications, Iran,Tehran, (2012).

[7] “Accident Insurance”, Summary Plan Description (2010) July, University of Chicago.

[8] Gaur, P.,” Neural Networks in Data Mining, International Journal of Electronics and Computer Science Engineering”, Vol 1, Iss 3, Pp 1449-1453, (2012).

[9] Chen, A., Pan, Y., Jiang, L., “Improving K-means Clustering Method in Fault Diagnosis based on SOM Network”, Journal of Networks, Vol 8, Iss 3, Pp 680-687, (2014).

[10] Khanchouch, I., Boujenfa, Kh., Limam, M., “An Improved MULTI-SOM Algorithm”, International Journal of Network Security & Its Applications, Vol 5, Iss 4, Pp 181-186 , (2013).

[11] Ming, M., Chiang, T., “Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads”, Journal of Classification 27 , (2009).

[12] Konstantinos Tsiptsis, A. C. “*Data Mining Techniques in CRM: Inside Customer Segmentation”.* Athens,: John Wiley & Sons, Ltd. (2012).

[13] Minaee, B., Nasiri, M., Hassani, D., Shenasa, E., “Step by step instruction Data Mining with Clementine”. Publications of the Engineering & Research Group(SAHER), Iran, Tehran, (2013).

[14] Dr. Chandra, E., Rajeswari, J. , “A Survey on Data Classification using Machine Learning Techniques” , International Journal of Engineering Science and Technology, Vol 3, Iss 10, Pp 7397-7401, (2011).

[15] Patil, N., Lathi, K., Chitre, V.,” Comparison of C5.0 & CART Classification algorithms using pruning technique”,

International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 4, June, (2014) .

[16] Zurada, J., Karwowski, W., & Marras, W. "A nural network-based system for classification of industrial jobs with respect to risk of law back disorders due to workplace design" .Ooccupational ergonomics, 49-58, (1997).

[17] Williamson, A., & Feyer, A.. "a classification system for causes of occupational accidents for use in preventive strategies". Scandinavian Journal of Work, Envirronment & health, 302-312, (2009).

[18] Deyin, H., Jing, Z., & Mao, L. "application of health risk classification method to assessing occupational hazard in china". 3rd International Conference on Digital Object Identifier, (pp. 1-5), (2009).

[19] Nearkson chau. Gerom C, G. "Contribution of occupational hazards and human factors in occupational injuries and their assiciations with job , age and type of injuries in railway workers". International Occupation Environment Health, 517-525, (2012).