ISBN 978-83-60810-14-9
ISSN 1896-7094

Improving Naïve Bayes Models of Insurance Risk by Unsupervised Classification

Anna Jurek
Institute of Mathematics
Technical University of Lodz
Wolczanska 215, 90-924 Lodz, Poland
Email: ankajurek@poczta.fm

Danuta Zakrzewska
Institute of Computer Science
Technical University of Lodz
Wolczanska 215, 90-924 Lodz, Poland
Email: dzakrz@ics.p.lodz.pl

***Abstract*—In the paper application of Naïve Bayes model, for evaluation of the risk connected with life insurance of customers, is considered. Clients are classified into groups of different insurance risk levels. There is proposed to improve the efficiency of classification by using cluster analysis in the preprocessing phase. Experiments showed that, however the percentage of correctly qualified instances is satisfactory in case of Naïve Bayes classification, but the use of cluster analysis and building separate models for different groups of clients improve significantly the accuracy of classification. Finally, there is discussed increasing of efficiency by using cluster validation techniques or tolerance threshold that enables obtaining clusters of very good quality.**

I. Introduction

RISK management is one of the most important area of interests in insurance industry. All the activities of insurance companies are usually connected with risk of policyholders claiming for big damages and the necessity of paying them a large amount of money. The level of the risk became the crucial factor influencing insurance company profits and may even decide on prosperity or failure in the industry. Evaluation of the risk level plays an important role in defining insurance companies' strategies. Having data of all the clients together with the historical information of damage claiming, insurers can estimate risk connected with the sale of every insurance policy. Building risk evaluation models may help them to make the decision concerning acceptance or rejection of new insurance applications. There exist different analysis methods that may support that process, as the most important there should be mentioned: predictive modeling, risk modeling, scoring or risk level qualification.

In the paper there is considered using of data mining techniques to classify life insurance applications as good or bad from the point of view of risk undertaken by the company. By examining historical data of insurance company, clients with certain characteristic features, may be qualified into the groups of different risk level. We investigate application of Naïve Bayes classifier for predicting the risk and for assigning every new customer into the proper group. Naïve Bayes algorithm is a very simple tool for machine learning and data mining. Despite of the fact that the conditional assumption, on which it is based, is rarely fulfilled in the real world, it ensures high accuracy of obtained results. The superb performance of its classification effects was explained in [1]. We propose to increase the efficiency of the classification by using clustering techniques in the preprocessing phase. The experiments showed that, however results obtained by building Naïve Bayes models, measured by the percentage of correctly qualified instances, were satisfactory, but application of clustering algorithm in the preprocessing stage improved significantly the accuracy of the classification. What is more using cluster validity measurement techniques enables choosing of optimal clustering schema and the further increase of correctness of classification.

The paper is organized as follows. In the next section some research concerning application of data mining algorithms in insurance industry is presented. In Section 3, data preparation process as well as applied algorithms are introduced. Then in Section 4 experiments and their results are described. Section 5 presents discussion and evaluation of obtained results. Finally some concluding remarks and future work are proposed.

II. Relevant Research

Analysis of large historical data sets has been used in actuarial investigations that concerns making optimal underwriting and price decisions for many years [2]. Development of data mining techniques enabled to increase an effectiveness of data analysis in insurance area. The main research focus on building predictive and risk evaluation models, that allow for identifying customers' behaviours and supporting insurance decision making. Especially data mining techniques were examined for fraud detection, discovering insurance risk or ameliorating customer services (see [3], [4], [5] for example). Examination of different application areas and their technical challenges are presented in [6]. In [7], author described, how the financial situation of the insurance company, may be improved by using such techniques as: decision trees, generalized linear modeling or logistic regression. In [5], there is considered effectiveness of association rules and neural segmentation for analyzing large data sets collected in health insurance information system. Authors considered detecting patterns in ordering pathology services as well as classifying practitionners, and they concluded that data mining techniques allow to obtain the results that are not achievable by conventional methods. Classification trees were examined in the Worker Compensation insurances area in [8], where CART algorithm [9] was used to build a model that classifies each claim as “likely to become more serious” or “not to become more serious”. Comparison of examined algorithm with logistic regression showed that CART model guarantees better classification results. In the same paper, there were also considered classification trees and hybrid modeling for building predictive models of hospital costs in health insurance. Decision trees were also used, in [10], to generate a predictive model, that may help insurance companies in the identification of claims which are the most likely to generate cost savings.

Data mining techniques were also investigated in fraud detection in health insurance area, taking into account such methods as k-Nearest Neighbor [11], multi-layer perceptron network [12] as well as heuristics and machine learning [13]. Broad review of application of neural networks, fuzzy logic and genetic algorithms in insurance industry can be found in [14]. An overview of fuzzy logic applications in insurance is presented in [15], where two approaches were studied: fuzzy logic applied separately and its combination with neural networks and genetic algorithms. The author reported application of fuzzy logic techniques in such insurance areas as classification, underwriting, projected liabilities, pricing, asset allocation, and investments. As the main fuzzy logic tool used in risk and claim classification, there was presented c-means clustering [15], which in turn was recognized, by Derrig and Ostaszewski, who classified insurance claim depending on level of fraud, as valuable addition to other methods but not the best technique (see [16] for example) to use separately.

Performance of Naïve Bayes models was mainly investigated in fraud detection area (see [17], [18]). Many empirical comparisons between Naïve Bayes technique and modern decision trees: C4.5 and C4.4 as well as Support Vector Machine, showed that Naïve Bayes predicts equally well [19]. In comparative study presented in [17], authors concluded that smoothed Naïve Bayes gave better results than C4.5 decision trees in automobile fraud insurance claims. Experimental study in individual disability income insurance fraud detection [20] showed that Naïve Bayes predictive models outperformed decision tree and Multiple Criteria Linear Programming models in terms of classification accuracy.

The accuracy of Naïve Bayes classifier was improved by combining it with other methods in several papers (see [21], [22], [23], [24] for example). Naïve Bayes classifier was enhanced by application of features selection methods, discretization or using boosting procedures. In [21], authors introduced Selective Bayesian Classifier (SBC), which uses only attributes not removed by C4.5 decision trees. They run C4.5 algorithm on the 10% samples of the all data shuffled together. That action was repeated five times, then the set of attributes was built as an union of those appearing only in the first three levels of the simplified decision trees. Created that way set of selected attributes was taken into account by Naïve Bayes classifier. Experiments conducted on the ten data sets showed that SBC learns faster than Naïve Bayes classifier on all the data sets, and the obtained results were improved up to 7.9%.

In [23], there were used together: dicretization, feature selection and boosting procedures, as techniques for Naïve Bayes improvements. Discretization of attributes with continuous values, was done by applying entropy-based method [25]. Features were selected by using filter that computes empirical mutual information between features and discard low-valued features, measured by gain ratio. The gain ratio value of selected attributes were to exceed a fixed threshold. In the proposed algorithm there was included boosting technique Adaboost introduced in [26]. During each iteration of the algorithm, there is applied entropy discretization technique and redundant attributes are removed by using the gain ratio feature selection method. Such defined algorithm is not as easily interpretable as simple Bayes model, but much more comprehensible than neural networks for example. Experiments conducted on 26 data sets showed that the proposed method was more accurate than Bayes algorithm in 12 cases, with the average error rate of about 20% less than that of simple Bayes algorithm. Another improvement of Naïve Bayes model, called hidden Naïve Bayes was proposed in [24]. The main idea of the proposed model consists in creating attribute hidden parents, that represent influence of all the other attributes. Experiments done on 36 data sets showed that considered algorithm outperforms Naïve Bayes, SBC as well as C4.5 decision trees.

III. Methodology

In considered model risk evaluation is based on classification rules, that may be built by exploring historical data collected in daily activity of insurance companies. Having characteristic features of all customers together with history of their policies and their claims, insurers can determine groups of different risk level. In our research, three groups of clients are distinguished. The first one, containing customers of low insurance risk (the best clients), the second one of medium risk level and the third group, that may be characterized by high risk (clients that should be avoided). The main idea consists in classifying each of new customers to one of the groups to predict the insurance risk. On the basis of information about the potential client, insurers can almost automatically estimate risk and refuse or accept potential customer application.

In our investigations two different approaches are considered: in the first one classification is based on all the data contained in the database; in the second one customers' data are divided into clusters of certain similarities and different classification rules are built for each segment separately. In that approach the new client is firstly assigned to one of the groups by unsupervised classification and then, the decision rule appriopriate to considered cluster is applied.

A. Data preparation

In the present study, we will limit our research into life insurance risk evaluation. However, insurance companies are very interested in finding effective risk evaluation models, but they are very sensitive and reluctant in sharing their data publicly, they do not allow for using their data even

TABLE I
ATTRIBUTES OF LIFE INSURANCE DATABASE

No	Attribute name	Attribute type
1	Sex	qualitative
2	Profession	qualitative
3	Region	qualitative
4	Hobby	qualitative
5	Drinking alcohol	qualitative
6	Smoking	qualitative
7	Disease	qualitative
8	Weight	quantitative
9	Age	quantitative
10	Blood pressure	quantitative
11	Maritial status	qualitative

for research goal. Problems connected with that fact, and its consequences were described with details in [10], where the author recognized lack of the access to real data as the main reason of limitations in understanding data mining and predictive modeling in insurance area.

We base our research on artificially generated data sets, with different attributes that may play a crucial role in profitability of life insurance. Considered client features, are strictly connected with medical exams that customers are obliged to fulfill. All the attributes, used in the research represent information required in medical exam of Swiss Life Insurance and Pension Company [27]. Some exemplary tests, containing similar questions, may be also found at [28]. All the attributes are presented in Table I. Most of them are of qualitative type, except of the following: “weight”, “age” and “blood pressure”.

For the purpose of Naïve Bayes classification all numerical data should be dicretized into ranges partitioned into intervals. In case of the attribute “weight” three ranges defined as underweight, proper weight and overweight, may be distinguished; in case of attribute “age” values may be binned into five ranges and for “blood pressure” we may have three: low, normal and high. Categorical data, in turn, should be binned into meta - classes (for example: region instead of city). This operation is necessary because Naïve Bayes model relies on calculating probability and cardinality of values should be reduced. Since all attributes are used in the classification process, all of them should be binned.

B. Naïve Bayes classifier

Naïve Bayes models are the simplest forms of Bayesian network for general probability estimation, detailed description of their functionality was presented in [29]. Naïve Bayes classifier, assumes conditional independence of input data. This is the strong assumption, that seems to be unrealistic in the insurance domain, but a history of empirical studies shows that even in such cases the method presented good performance [18].

By application of Naïve Bayesian algorithm, we obtain probability distribution of belonging into classes. In uncertain cases objects may not be assigned into any group (reject option) or similarly to fuzzy logic techniques may be allocated into more than one class. Other advantages of classifier probabilistic output, such as changing utility functions, compensating for class imbalance or combining models, were described in [30]. Naïve Bayesian algorithm also naturally deals with missing values, what is difficult to achieve by decision trees or neural networks methods. What is more, obtained models are very easy to understand, without further investigations, that feature is not valid for all methods, to mention neural networks as an example. Comparisons of the performance of Naïve Bayes classifier and other classification techniques, like decision trees, neural networks, kNN, Support Vector Machine or rule-learners, were presented in [31]. As main features, which allow Naïve Bayes technique to outperform other algorithms, there were mentioned: speed of learning and classification, tolerance to missing values, explanation ability as well as model parameter handling. To increase of the classification accuracy, the author suggested application of ensemble methods [31].

Naïve Bayes model is based on maximum likelihood, that uses very well known Bayes’ formula:

$$P(H_j/A) = \frac{P(A/H_j)P(H_j)}{\sum_{i=1}^{n} P(A/H_i)P(H_i)}, \tag{1}$$

where $j \in 1...n$,$P(A/H_j)$ means conditional probability and is defined as:

$$P(A/H_j) = \frac{P(A \cap H_j)}{P(H_j)} \tag{2}$$

H_j is an event, that means belonging to the group of risk. A is a vector of customer attributes. $P(H_j/A)$ is the probability that person described by A belongs to H_j; $P(H_j)$ means probability of belonging to group H_j. $P(A/H_j)$ is the probability that customer from H_j is described by A.

Algorithm compares features’ vector of a new customer with all records in the database and computes the probability of memberships of all of the groups, then the considered client is assigned into the group of risk, for which probability of belonging is the highest. Efficiency of the algorithm may be evaluated by checking assignements for test set of customers’ data, by comparing of obtained results with real belongings to the group of certain risk level. Assessment of classification accuracy is done by calculating the percentage of correctly classified records. Addidtional advantage of the technique is that it does not have to use all attributes and can work just with few selected ones. The choice of the attributes that guarantee the most accurate results may be done by the use of training data sets in the preprocessing phase.

C. Clustering

Cluster analysis techniques become very popular in customer segmentation area. One of the main advantage of the clustering technique is that it does not assume any specific distribution on the data. The main disadvantage of the method is the high dependence of experts’ opinions in many cases. There exist many clustering techniques that may be used in customer grouping, the broad review of the most popular of them is presented in [32]. In the current research, k-means algorithm has been chosen, because of its simplicity and efficiency. This algorithm performed very well in experiments concerning credit risk scoring [33]. However the method depends significantly on the initial assignments, what may entail in not finding the most optimal cluster allocation at the end of the process, but k-means is very efficient for large multidimensional data sets [32].

Segmentation is done according to attributes that may play the most crucial role in life insurance: “Drinking alcohol”, “Smoking”, “Profession”, for different number of required clusters. The distance between two objects is measured by Manhattan function. As optimal clustering schema may differ depending on data sets, validity measurement techniques are used to evaluate obtained clusters and to make the best choice for their numbers.

Different cluster validity indices were introduced to examine quality of grouping results. All of them and their application for different clustering algorithms are broadly investigated in [34]. To find out the best clustering schema, it has been chosen Davies-Bouldin (DB) index because of its simplicty. It is based on dispersion and cluster distance measures, which means taking into account: internal variance and external similarity. Low DB value means that clusters are compact and well separated. DB is defined as follows [34]:

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_i, \tag{3}$$

where k is the number of clusters, R_i is the ratio of dispersion measures and distance of clusters C_i and C_j: $R_i = max\left\{R_{ij}, i \neq j\right\}, R_{ij} = (S_i + S_j)/D_{ij}$. D_{ij} is measured by the distance between centroids ν_i and ν_j of C_i and C_j respectively, and dispersion measure of cluster C_i is defined by:

$$S_i = \left(\frac{1}{|C_i|}\sum_{x \in C_i} d^p(x, \nu_i)\right)^{1/p}, p > 0, \tag{4}$$

where $d(x, y)$ is the distance between x and y.

IV. Experiments

Experiments were done on artificially generated data sets, with attributes presented in Table I. Two databases were filled with values randomly. The first one contains five hundred records with eight attributes: “sex”, “profession”, “region”, “hobby”, “drinking alcohol”, “smoking”, “disease” and “weight”. The second one, of eight hundred records is characterized by eleven attributes, all of them are presented in Table I. The main aim of the experiments was to examine how the use of clustering techniques in the preprocessing stage may influence classification accuracy. There were compared results of classification done by using all the records of considered data sets, with effects obtained on segments of similar customers. It was also examined how the quality of results changes depending on choice of attributes taken into account during classification in different clusters.

Considered databases contain records that characterize people who bought life insurance policies. All the customers, according to their characteristic features were assigned into one of three groups: of high risk—the ones who cause the biggest financial losses for the company; of low risk—the ones who are expected to be rather profitable than harmful; of medium risk—the ones that do not belong neither to first nor to the second group. Finally there are three groups of customers with different level of risk. During experiments, all records from databases were divided into training and test sets. Records of the first ones took part in classification process, while records of the second ones were to check the classification accuracy, by comparing obtained results with real assignments into the groups of risk.

Experiments were divided into three stages. In the first one classification was made on the entire datasets, during the second stage groups of similar customers were found by clustering, and classification was done on each cluster separately. Finally the role of the attributes in obtaining high accuracy is examined.

A. Classification on entire data sets

Results of the classification process, made on entire data sets, were examined taking into account different combinations of attributes. At the beginning, all the attributes were used. The results were rather poor, only 57% of customers were classified correctly in case of the first database (8 attributes)and 59% for the second database (11 attributes) . Much better effects were achieved taking into consideration different combinations of parameters. For the first database, the best results of 71% of objects classified correctly were obtained for combination of 5 from among 8 attributes: “weight”, “smoking”, “profession”, “region”, “disease”. For the second database, also combination of 5 from among 11 attributes: “drinking alcohol”, “smoking”, “disease”, “blood pressure”, “age”; gave the best results of 76% correctly assigned clients.

The main reason for such an improvement in case of usage of selected attributes is the fact that Naïve Bayes algorithm, contrarily to decision trees, cannot recognize which of features have more influence in the classification process. One attribute cannot affect membership of the object in any of the groups. During experiments, it was noticed that including some of the attributes into the classification process can even decrease its efficiency. If number of attributes is not so big, like in considered cases, there may be tried different combinations of attributes to choose the best ones. Another possibility is to ask experts.

B. Classification on clusters

In this part of the experiments, all customers were grouped by unsupervised classification, before applying Naïve Bayes models. To maintain the proper balance between the weight of each customer features, different attributes were taken into account in every stage of the classification. During the

TABLE II
CLASSIFICATION ACURACY FOR ALL ATTRIBUTES AND 3 CLUSTERS

Data set	Correctly classified
Cluster 1	51.35%
Cluster 2	65.21%
Cluster 3	71.5%
All data	63%

TABLE III
CLASSIFICATION ACURACY FOR ALL ATTRIBUTES AND 2 CLUSTERS

Data set	Correctly classified
Cluster 1	63.33%
Cluster 2	75%
All data	68%

TABLE VI
CLASSIFICATION ACURACY FOR ALL ATTRIBUTES AND 3 CLUSTERS

Data set	Correctly classified
Cluster 1	79.31%
Cluster 2	100%
Cluster 3	65%
All data	74%

TABLE VII
CLASSIFICATION ACURACY FOR ALL ATTRIBUTES AND 2 CLUSTERS

Data set	Correctly classified
Cluster 1	73.33%
Cluster 2	62.5%
All data	69%

investigations, there were examined the influence of different combinations of attributes as well as the required number of clusters on the accuracy of obtained results. For both data sets, clustering was done according to three customer features, while the other attributes were used to classify clients into appriopriate groups of risk level. There were considered two different approaches: using the same attributes for all the clusters or distinguishing the choice of attributes depending on clusters.

Some exemplary results for the first database (8 attributes, 500 instances), where clustering was done according to the attributes: “drinking alcohol”, “smoking”, “profession”, for different numbers of clusters, are presented in Table II and Table III. In both of the cases the set of 5 other attributes is used for Naïve Bayes classification.

Table IV and Table V shows percentage of correctly nested instances, when classification rules are built by using different attributes for each cluster. The best choice of attributes, measured by the highest accuracy of classification are presented in the last columns.

Comparison of values presented in Table II, Table III, Table IV and Table V shows that the quality of obtained results is better when different attributes are used to build models for each cluster. For example, if the required number of clusters is equal to 3, total accuracy is increased of 17 percentage points. In the case of two clusters the increase was of 10 percentage points. It can be easily noticed that using of two and three clusters gives very similar effects. For three segments, 2 attributes were used for building classification models in each cluster, while for two segments, in both of them, three of five attributes were used. Results presented in the tables were the best from among those obtained in different experiments. In real databases, with greater number of attributes, the opinion of an expert concerning the feature selection may be very useful.

Exemplary results for the second database (11 attributes, 800 instances) are presented in Table VI and Table VII. Similarly to the first database, customers are segmented according to 3 of the following features: “drinking alcohol”, “smoking”, “profession”, with the required number of clusters equal to 2 or 3 and classification models built on the basis of 8 other attributes.

Results for decision rules built on different attributes in each cluster are presented in Table VIII and Table IX. Also in that case, classification effects obtained by using different attributes in each cluster are better. In case of the segmentation into

TABLE IV
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 3 CLUSTERS

Data set	Correctly classified	Attributes
Cluster 1	72.97%	weight, disease
Cluster 2	78.26%	weight, disease
Cluster 3	87.5%	weight, disease
All data	80%	

TABLE V
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 2 CLUSTERS

Data set	Correctly classified	Attributes
Cluster 1	75%	weight, disease, hobby
Cluster 2	82.2%	weight, disease, region
All data	78%	

TABLE VIII
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 3 CLUSTERS

Data set	Correctly classified	Attributes
Cluster 1	84.48%	disease, blood pressure, age
Cluster 2	100%	disease, blood pressure, age
Cluster 3	77.5%	hobby, disease, age
All data	82%	

TABLE IX
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 2 CLUSTERS

Data set	Correctly classified	Attributes
Cluster 1	81.66%	weight, disease, age
Cluster 2	77.5%	hobby, disease, age
All data	80%	

TABLE X
CLASSIFICATION ACURACY FOR ALL ATTRIBUTES AND 2 CLUSTERS

Data set	Correctly classified
Cluster 1	55%
Cluster 2	62.5%
All data	58%

TABLE XI
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 2 CLUSTERS

Data set	Correctly classified	Attributes
Cluster 1	66.66%	weight, drinking alcohol, hobby
Cluster 2	70%	weight, drinking alcohol, hobby
All data	68%	

two or three clusters, the situation is the same: classification models built for each cluster separately are less complex than the ones obtained for all the data. In all the cases only 3 attributes were used to build models for each cluster. Values of correctly classified instances presented in Table VIII and Table IX, are significantly higher than the ones presented in Table VI and Table VII, where classifications were done on the basis of all 8 attributes.

C. The choice of attributes

The aim of these investigations, was to check if for all the sets of attributes, segmenting of customers guarantees the improvement of classification results. After several experiments there were indicated attributes: “smoking”, “region”, “disease”, for which results are worse. Table X and Table XI present the effects of classification for the first data set (8 attributes, 500 instances), for clusters built according to those attributes.

It can be easily seen that the results of classification presented in Table X and Table XI are definitely worse than the ones obtained for the entire data set. The accuracy effectively increased after building classification models on different attributes, but it is still worse than the one obtained without using clusters.

V. EVALUATION

It was shown in the previous section that classification accuracy depends on the choice of attributes as well as the number of clusters. It may be expected that the results of classification should be more accurate for clusters of better quality. Well constructed clusters should have high internal and low external similarity. Calculating of DB index defined by (3) allows for choosing the best schema (optimal number of clusters) that will guarantee good quality of obtained clusters. It let us avoid experimental choice of required number of clusters. In considered cases, for the first database Davies Bouldin index is respectively equal to 0.0539 for two clusters and 0.0545 in case of three clusters. Comparisons of Table II and Table III as well as Table IV and Table V show that the effects in both cases are very similar, what is reflected in proximity of DB values. In the case of second database, DB value for two clusters is equal to 0.0547, while for three clusters to 0.0534, which should guarantee the better quality of clusters. Indeed, comparing Table VII and Table VIII as well as Table IX and Table X, it can be easily noticed that the choice of three clusters gives better effects in both cases.

TABLE XII
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 3 CLUSTERS, AFTER APPLICATION OF THRESHOLD IN EACH CLUSTER

Data set	Correctly classified	Attributes
Cluster 1	72.97%	weight, disease
Cluster 2	69.56%	weight, disease
Cluster 3	90%	weight, disease
All data	79%	

TABLE XIII
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 2 CLUSTERS AFTER APPLICATION OF TOLERANT TRESHOLD IN EACH CLUSTER

Data set	Correctly classified	Attributes
Cluster 1	83.33%	weight, disease, age
Cluster 2	60%	hobby, disease, age
All data	74%	

The next issue that should be taken into account is a presence of exceptions in data sets. During the segmentation process, such objects, that do not fit into any of the groups, usually are allocated into one of them. Existence of such elements may decrease quality of clusters and efficiency of the classification process. That situation may be avoided by removing the most distant objects from the clusters. The establishing of tolerance thresholds for each cluster may allow to isolate outliers and not to take them into account during classification process. In the current research as the threshold value for each cluster, the maximum distance between two objects in the generation, divided by the number of clusters, is used.

Table XII presents classification results in case of three clusters, for the first database, after application of tolerance threshold. Comparing Table XII and Table IV we can see that accuracy in Cluster 1 did not change, while in Cluster 3, it efficiently increased from 87.5% to 90%. In Cluster 2, however, we can observe a decrease from 78.26 to 69.56%. Similar situation may be noticed in the case of the second database and segmentation into two clusters. The results are presented in Table XIII

Comparison of Table IX and Table XIII shows that efficiency of classification increased in the first cluster but considerably decreased in the second one. It means that different value of tolerance threshold established for different clusters may give better results. If in the case presented in Table XIII we will use tolerance threshold only in the first cluster, we will obtain results with a general increase in accuracy, what can be seen in Table XIV.

TABLE XIV
CLASSIFICATION ACURACY FOR DIFFERENT ATTRIBUTES AND 2 CLUSTERS AFTER APPLICATION OF THRESHOLD IN FIRST CLUSTER

Data set	**Correctly classified**	**Attributes**
Cluster 1	83.33%	weight, disease, age
Cluster 2	77.5%	hobby, disease, age
All data	81%	

VI. CONCLUSION

In the paper, there is considered application of classification techniques for insurance risk evaluation. The idea is based on dividing clients into three groups of a different risk level. There has been chosen Naïve Bayes model as a classifier. Cluster analysis technique is proposed to improve classification accuracy. Experiments conducted on two data sets, characterised by different attributes of different number of records showed that building classification models for each cluster separately, ameliorates the accuracy of obtained results. For the first database number of correctly classified instances increased from 71% in the case of building the same classification model for all the data to 80%, in the case of differentiating classification models according to clusters, while for the second dataset the growth was from 76% to 82%.

The investigations showed that different results may be received for different number of clusters. To avoid determining a number of clusters experimentally, validation technique, which will indicate optimal schema of clusters of the best quality, may be applied. Classification accuracy can be increased by application of restrictions concerning objects in each cluster. However, experiments showed that establishing of a tolerant threshold, may also bring opposite effects. There should be worked out the strategy, that will allow to differentiate threshold values depending on clusters' properties. The experiments proved that classification process may give much better results by combining Naïve Bayes models with cluster analysis. Taking into account the data that insurance companies possess they may easily segment their customers and build risk models for each of the group separately. The experiments were conducted on artificially generated data sets, in the next step obtained results should be verified on the real data of insurance company.

Future research should also consist of development of the proposed model, by using for example SBC instead of Naïve Bayes classification, as the most crucial problem and the big challenge for researchers, at the same time, is the choice of the optimal combination of features for building classification models.

REFERENCES

[1] H. Zhang, "The optimality of Naïve Bayes,"*in the 17th FLAIRS Conference,* Florida, 2004.

[2] S. A. Klugman, H. H. Panjer and G. E. Willmot, *Loss Models: From Data to Decision,* John Wiley & Sons, New York, 1998.

[3] C. Apte, E. Grossman, E. Pednault, B. Rosen, F. Tipu and B. White, "Probabilistic estimation based data mining for discovering insurance risks,"*IEEE Intelligent Syst.,* vol. 14, 1999, pp. 49–58.

[4] J.-U. Kietz, U. Reimer and M. Staudt, "Mining insurance data," *in the 23rd VLDB Conference,* Athens, Greece, 1997.

[5] M. S. Viveros, J. P. Nearhos and M. J. Rothman, "Applying data mining techniques to a health insurance information system," *in the 22nd International Conference on Very Large Data Bases,* Bombay, India, 1996, pp. 286–294.

[6] S. J. Hong and S. M. Weiss, "Advances in predictive models for data mining," *Pattern Recogn. Lett.,* vol. 22, 2001, pp. 55–61.

[7] R. Mosley, "The use of predictive modeling in the insurance industry," *PINNACLE Actuarial Resources, INC.,* January, 2005.

[8] I. Kolyshkina and R. Brookes,"Data mining approaches to modeling insurance risk," *Report,* PriceWaterhouseCoopers, 2002.

[9] J. Han and M. Kamber, *Data Mining: Concepts and Techniques. Second Edition.* Morgan Kaufmann Publishers, San Francisco CA; 2006.

[10] S. P. D'Arcy, "Predictive modeling in automobile insurance: a preliminary analysis,"*in World Risk and Insurance Economics Congress,* Salt Lake City, Utah, August 2005.

[11] H. He, W. Graco and X. Yao, "Application of genetic algorithms and k-nearest neighbour in medical fraud detection," *In Proceedings of SEAL 1998,* Canberra, Australia, 1999, pp. 74–81.

[12] H. He, J. Wang, W. Graco and S. Hawkins, "Application of neural networks to detection of medical fraud," *Expert Syst. Appl.,* vol. 13, 1997, pp. 329–336.

[13] J. Major and D. Riedinger, "EFD:A hybrid knowledge/statistical -based system for the detection of fraud," *J. Risk Insur.,* vol. 69, 2002, pp. 309–324.

[14] A. F. Shapiro, "The merging neural networks, fuzzy logic and genetic algorithms," *Insur. Math. Econ.,* vol. 31, 2002, pp. 115–131.

[15] A. F. Shapiro, "Fuzzy logic in insurance," *Insur. Math. Econ.,* vol. 35, 2004, pp. 399–424.

[16] R. A. Derrig, K. M. Ostaszewski, "Fuzzy techniques of pattern recognition in risk and claim classification," *J. Risk Insur.,* vol. 62, 1995, pp. 447–482.

[17] S. Viaene, R. A. Derrig, B. Baesens and G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," *J. Risk Insur.,* vol. 69, 2002, pp. 373–421.

[18] S. Viaene, R. A. Derrig and G. Dedene, "A case study of applying boosting Naïve Bayes to claim fraud diagnosis,",*IEEE T. Knowl. Data En.,* vol. 16, 2004, pp. 612–620.

[19] J. Huang, J. Lu, Ch. X. Ling, "Comparing Naïve Bayes, decision trees and SVM with AUC and accuracy," *in Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* Melbourne, Florida, USA 2003.

[20] Y. Peng, G. Kou, A. Sabatka, J. Matza, Z. Chen, D. Khazanchi and Y. Shi, "Application of classification methods to individual disability income insurance fraud detection," *in ICCS2007. LNCS 4489,* Y. Shi, G. D. van Albada, J. Dongarra, P. Sloot, Eds., Springer, Berlin Heidelberg, 2007, pp. 852–858.

[21] Ch. Ratanamahatana, D. Gunopulos, "Scaling up the Naïve Bayesian classifier: using decision trees for features selection," *in Proceedings of Workshop on Data Cleaning and Preprocessing (DCAP 2002), at IEEE International Conference on Data Mining (ICDM'02),* Maebashi, Japan, 2002.

[22] P. Langley, S. Sage, "Induction of selective Bayesian classifiers," *in Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence,* Seattle, 1994, pp. 399–406.

[23] S. B. Kotsiantis, P. E. Pintelas, "Increasing the classification accuracy of simple Bayesian classifier," *in AIMSA2004. LNAI 3192,* C. Bussler, D. Fensel, Eds., Springer, Berlin Heidelberg, 2004, pp. 198–207.

[24] H. Zhang, L. Jiang, J. Su, "Hidden Naïve Bayes," *in Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI2005),* AAAIPress, 2004, pp. 919–924.

[25] J. Doughtery, R. Kohavi, M. Shami, "Supervised and unsupervised discretization of continuous features," *in Machine Learning: Proceedings of the Twelth International Conference,* A. Prieditis, S. Russell, Eds., Morgan Kaufmann Publishers, 1995, pp. 194–202.

[26] Y. Freund, R. E. Schapiro, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence,* vol. 14, 1999, pp. 771–780.

[27] Extended Medical Examination, $http : //www.swisslife.ch/etc/slml/slch/obedl/1/200/$ $316.File.tmp/form_9 15035_g esundheitspruefung.pdf.$

[28] $http : //personalinsure.about.com/od/life/a/aa112805a.htm$

[29] D. Lowd, P. Domingos, “Naive Bayes models for probability estimation,” *in Proceedings of 22nd International Conference on Machine Learning,* Bonn, Germany, 2005.

[30] K. P. Murphy, “Naive Bayes classifiers,” $http://www.cs.ubc.ca/\ murphyk/Teaching/CS340-Fall06/reading/NB.pdf$.

[31] S. B. Kotsiantis, “Supervised machine learning: a review of classification,” *Informatica,* vol. 31, 2007, pp. 249–268.

[32] M. N. Murty, P. J. Flynn and A. K. Jain, “Data clustering: a review,” *ACM Comput. Surv.,* vol. 31, 1999, pp. 264–323.

[33] D. Zakrzewska, “On integrating unsupervised and supervised classification for credit risk evaluation,” *Information Technology and Control,* vol. 36, 2007, pp. 98–102.

[34] G. Gan, Cha. Ma and J. Wu, *Data Clustering: Theory, Algorithms and Applications,* ASA-SIAM Series on Statistics and Applied Probability, SIAM: Philadelphia, ASA: Alexandria, 2007.