Analyzing Life Insurance Data with Different Classification Techniques for Customers' Behavior Analysis

Md. Saidur Rahman, Kazi Zawad Arefin, Saqif Masud, Shahida Sultana and Rashedur M. Rahman

Abstract Analyzing data of life insurance companies gives an important insight on how the customers are reacting to the offered insurance policies by the companies. This information can be used to predict the behavior of future policy holders. Life insurance companies maintain a large database on their customers and policy related information. Data mining technique applied with proper preprocessing of data prove to be very efficient in extracting hidden information from data stored by life insurance companies. There are many data mining algorithms that can be applied to this huge set of data. The main focus of our work is to apply different classification techniques on the data provided by a life insurance company of Bangladesh. Attribute selection techniques are applied to properly classify the data. Classification techniques proved to be very useful in classifying customers according to their attributes. A comparative analysis of the performance of the classifiers is also reported in this research.

Keywords Data mining · Data balancing · Life insurance · Machine learning · Custer behavior analysis

Md. Saidur Rahman · K.Z. Arefin · S. Masud · S. Sultana · R.M. Rahman (✉)
Department of Electrical and Computer Engineering, North South University,
Plot – 15, Block – B, Bashundhara, Dhaka 1229, Bangladesh
e-mail: rashedur.rahman@northsouth.edu

Md. Saidur Rahman
e-mail: saidurrahman@northsouth.edu

K.Z. Arefin
e-mail: kazi.arefin@northsouth.edu

S. Masud
e-mail: saqif.masud@northsouth.edu

S. Sultana
e-mail: shahida.sultana@northsouth.edu

© Springer International Publishing AG 2017

D. Król et al. (eds.), *Advanced Topics in Intelligent Information and Database Systems*, Studies in Computational Intelligence 710,
DOI 10.1007/978-3-319-56660-3_2

1 Introduction

Life insurance company deals with a huge amount of data acquired from their policy holders. The data is generally related with customers' personal information, health information and current financial status. Companies also record information about the policy the customers are buying and premium payment. Most often these attributes indicate whether a customer can be regular or irregular in his installment payment. This information is very crucial for an organization like life insurance companies because they want to design their policy in such a way that it can attract most of the customers who are willing to pay installments.

A life insurance company is a financial organization which offers insurance policies to its customers. An insurance policy is a legal agreement between an insurance provider and an insurance policy holder. The insurance company promises to pay a designated amount of money to the insurance holder according to the insurance policy. The amount of money and conditions in which the policy holder is paid depends on the insurance policy or agreement offered by the insurance company. Policy holders often have to pay a monthly or yearly premium against their policy. Here comes the matter of regularity of a customer. The insurance company is interested about the regular customers because it increases their chance of profit as a company. Again customers who are regular are benefited from the insurance policy they have bought because it ensures their chances of successful insurance claim.

To classify the customers as regular or irregular different data mining techniques prove to be useful. Mainly classification algorithms are used to classify different types or classes of data from a dataset. Our target is to classify the customers based on their given attributes so that we can predict the class label for future customers. Classification techniques such as JRIP, Naïve Bayes, IBK, PART etc. are used and their performance is compared to find the best suited classifier for the acquired dataset. Data mining techniques are generally used to find interesting patterns in the dataset to provide useful information for the future. For our dataset the data is imbalanced on a particular class label. So we have to balance our data using balancing techniques such as Random Over Sampling (ROS) and Random Under Sampling (RUS).

2 Related Work

Kirlidog and Asuk's [1] worked on fraud detection approach on health insurance company data. They applied anomaly detection, clustering techniques and classification to detect anomaly in data. First they used statistical approach on the data collected for calculating percentage of rejected claims. Next they divided their anomaly detection into two criteria. First they considered excessive claim by different types of health centers. Next they considered individual health centers and calculated excessive claim for them. The combined result gave them a predictive model for anomaly detection

Xiayun and Danyue [2] also designed an algorithm called RB algorithm to detect outlier in the evaluation of client moral risk. The algorithm along with the application of density factor detected the chances of client to have high moral risks. This RB algorithm first selected an initial resolution for a data set and the density was updated if the cluster size increased with the changing resolution of the algorithm. They also considered seven paramount attributes that effected greatly in client moral risk. They worked on a data of Chinese Health Insurance Company.

Yan and Xie [3] worked with data mining techniques on the Chinese insurance companies. They proposed decision tree based classification and application of data mining in CRM (Client Relation Management) model. They also proposed to apply it in risk management. They also mentioned about the need of insurance companies to maintain a large data center or warehouse to efficiently store the information.

Goonetilleke and Caldera [4] used data mining techniques on life insurance company data to evaluate the possibility of attrition of a customer. They first selected effective attributes using CFS (co-relation based feature selection) method. Then they used different classification algorithms and ranked the effective attributes. As the initial data set was imbalanced they used cost sensitive learning approach. They considered different stages of an insurance policy and tried to evaluate the attrition probability of customer depending on the current stage of his insurance policy. They also evaluated the classification techniques on different classification evaluation metrics. Finally they developed a cost matrix to evaluate the attrition cost of a customer.

Thakur and Sing [5] used decision tree based classification technique for developing a prediction system on customer data. They had a training data of customers who wanted vehicle insurance from online. Based on the customers' attributes they classified new customer for their interest in online insurance. They evaluated their classifier based on its accuracy and error rate. Their main target was to classify based on the age of customer and educational status and the type of vehicle they own. They built a system that provided all the necessary information for vehicle insurance online.

3 Dataset and Tools

We have collected the data from Prime Islami Life Insurance Company Ltd., of Bangladesh which has the data from almost every division of Bangladesh. The timeline of data was from 2011 to 2014 and it has data of about 282,282 policy holders. We want to determine that if a customer is regular or irregular in installment payment. So we introduced a new attribute named 'Regularity'. This attribute consists of two values which are regular/irregular. We have also assumed that the customers who are regular can complete the installments of policy and the irregular ones will fail to do it. To properly assess a customer as capable of full installment payment we required data that would range over a larger timeline which is rare to acquire so we assumed the regularity factor. We determined the regularity based on the following facts which are the starting date of a policy, the last payment dates of every policy and if a customer has paid his dues before these dates. One major factor is that the dataset was greatly imbalanced considering the 'Regularity' attribute. We had to maintain the ratio throughout data preprocessing.

We used "WEKA" for implementing different algorithms of balancing and classification techniques and used MySQL server for storing and preprocessing the dataset.

4 Methodology

Our approach contains mainly three steps—"Data Preprocessing", "Attribute Evaluation" and "Classification Technique Implementation". At first in "Data Preprocessing", we prune the less important attributes from the dataset and then prune the incomplete records. Sampling is used as the original dataset is huge. But we need to maintain the ratio of 'class' attribute in sample dataset as of that in original dataset. Then, in "Attribute Evaluation", we figure out which attributes are worthy to consider for classification. In "Classification Technique Implementation", we explicitly balance the dataset and then apply different classification techniques.

To find out the most effective classifier, we go through a number of processes. We can breakdown our tasks or processes in following parts

1. Data Preprocessing
2. Attribute Evaluation/Ranking
3. Classification Technique Implementation

   3.1 Class Label Balancing
   3.2 Classification

To preprocess the data it is required that we understand the collected data properly. As insurance policies are difficult to understand in general we have to study about the methods of the storage of data by an insurance company. We prune the data so that the attributes that could have maximum effect on specifying a customer as regular or irregular could be obtained. In the attribute evaluation/ranking section we apply certain attribute evaluator techniques and compare them so that the best set of attributes could be chosen for classification. We again apply class label balancing techniques so that we could balance the data that is previously unbalanced with respect to class label. We use a number of classification techniques such as JRIP, SMO etc. and compare them so that we could determine the best classification model. Details of these parts are described in the following subsections.

4.1 Data Preprocessing

Data preprocessing part is a two steps process—"Pruning" and "Sampling".

4.1.1 Pruning

The initial data set contains more than 100 attributes. Most of these attributes are found to be irrelevant by us. Based on our study and research we select 10 attributes (POLICY TERM, AGE, SEX, OCCUPATION, URBAN-RURAL, MARITAL STATUS, SUM-ASSURED, DIVISION, PREMIUM PAYMENT MODE and REGULARITY) that we figure out effective in determining the regularity of a customer. We also have to discard many of the records because they contain null values. Brief descriptions of these attributes are as follows.

(a) Policy Term (PT): This refers to the time span through which the policy holder will pay his or her premiums.
(b) Age (A): It points the age of the policy holder when s/he starts the policy.
(c) SEX (SX): It marks the gender of the policy holder.
(d) OCCUPATION (O): The occupational status (i.e., teacher, electrician or housewife) of the policy holder, when s/he starts the policy, is described here.
(e) URBAN-RURAL (UR): This refers to the policy holders dwelling status whether s/he is a city dweller or lives in village area.
(f) MARITAL STATUS (MS): It marks the marital status (married, unmarried or divorced, single) of the policy holder.
(g) SUM-ASSURED (SA): The pre-decided amount which the insurer promises to pay the nominee in case of the policyholder's death.
(h) DIVISION (D): It marks the division (Dhaka, Sylhet etc.) of Bangladesh, the policy holder registered his or her policy from.
(i) PREMIUM PAYMENT MODE (P): It marks the breakdown of the payment of the policy holder's total premium (i.e., yearly or monthly)
(j) REGULARITY (R): It marks the policy holder as "regular" or "irregular" based on whether s/he is paying the premiums on time.

4.1.2 Sampling

We initially have a data set of about 282,282 policy holders and we prune the data to about 10,000 policy holders for faster processing of data. We also maintain initial ratio of the data so that the integrity of the data is not lost while pruning. The reason behind doing so is that the use of classification techniques over full dataset is highly time-consuming and also it does not make any notable accuracy.

4.2 Attribute Evaluation

The attributes selected during pruning as relevant for classification often prove to be ineffective in action. To analyze effectively, we may have to discard some attributes which are selected during pruning. If we take all the 10 attributes for classification the classification would be unnecessarily sparse and difficult to interpret. The model also tends to be data driven which is it tries to memorize its training set and fits the testing set accordingly.

We apply different attribute evaluator techniques to determine the most effective attributes for classification. The first two techniques which we apply are 'information gain attribute evaluation' and 'gain ratio attribute evaluation techniques'. From 'information gain attributes evaluation' we found P, PT, SA, A, O and from 'gain ratio attribute evaluation' we found P, PT, SA, A, MS attributes are relevant.

We use "Greedy Step Wise" algorithm as the search method for Correlation based Feature Selection (CFS). In our approach, we get "Premium Payment Mode" (P) as the most effective attribute for our dataset.

We also use 'Classifier attribute evaluation' technique which determines the effective attributes for the designated classifier.

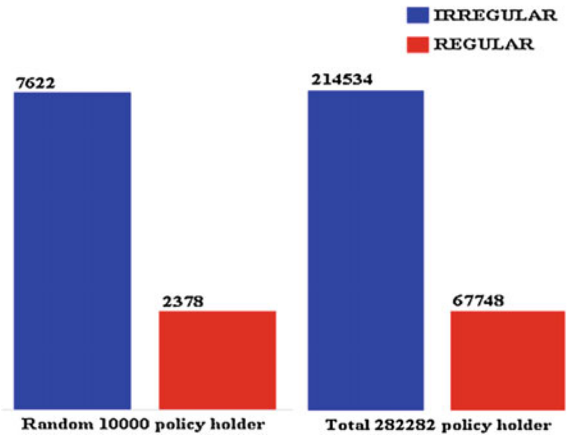
For each classifier that we use later we chose the "worthy" attributes by applying these techniques. Here, the "worthy" attributes are those which are common in the findings from the 'information gain attribute evaluation' and 'gain ratio attribute evaluation techniques' and 'Classifier attribute evaluation' for that specific classifier. This is the common practice for finding the relevant attributes for classification.

4.2.1 Classification Technique Implementation: Class Label Balancing

In Fig. 1, we see that the original ratio of regular and irregular percentage is maintained in the sample data. Maximum (almost 76%) of our data belongs to 'irregular' class which indicates there is an imbalance in the data set. The problem occurs when we try to apply classification techniques on this imbalanced data set as the result of the classification becomes biased towards the class of larger percentage. It means that most of the prediction is labeled as the major class which in this case is the 'irregular' class. To balance our data set we decide to apply both under-sampling and over-sampling techniques. We apply 'RUS' (Random Under Sampling) for under-sampling and SMOTE [6] (Synthetic Minority Over Sampling Technique) for over-sampling.

RUS algorithm tries to minimize the number of majority class randomly by a ratio. In our analysis we set the distribution spread value as 1 so that the ratio of majority and minority becomes 1. SMOTE algorithm tries to over sample the minority class using k-nearest neighbor calculation without sample replacement. It means SMOTE creates synthetic values by calculating random k-nearest neighbors from the designated minority class and increases the number of tuples in the data set. We set the parameters of the SMOTE algorithm as follows, K = 5; It indicates

Fig. 1 Regularity ratio in the main dataset and sample dataset



the number k-nearest neighbors considered T = 2378; The original number of records in data set labeled as regular N = 220.5%; amount of SMOTE percentage which means the minority class is oversampled by 220.5% from the original data. Total 5244 data of regular class is synthetically created and added to our data set.

4.2.2 Classification Technique Implementation: Classification

After pruning, balancing and attribute selection of data, we apply different classification techniques to figure out which techniques are most effective and yield more accurate results. We apply two methods to balance our dataset. One of them is the RUS and another one is SMOTE which balances the data by under sampling and over sampling respectively. We applied RIPPER, Naïve Bayes, IBK, SMO, Multilayer Perceptron and PART on our data. Each algorithm is tested on both balanced data acquired from applying the RUS and SMOTE methods. We have compared the ROC curve among the two balanced data for each algorithm. This process is done to figure out which algorithm is effective. Below some classification techniques are briefly discussed.

'Ripper' is a classification algorithm. It is a rule based direct classifier. It is best for imbalanced class distribution as it uses a validation set for preventing over-fitting of model. It uses general to specific rule growing approach with Foil's information gain to measure the performance of the rule. It works well with large data set as it does not consider all of the training data for rule production. 'Naïve Bayes' is another classification approach which uses the Bayesian probability considering every attribute independent from each other and calculates their probability when probability of class is given. It chooses the value of class label that maximizes the output of probability of attributes used for classification. This algorithm considers the attributes to be independent. ROC curve for the implementation of JRIP and Naïve Bayes are shown in Fig. 2.

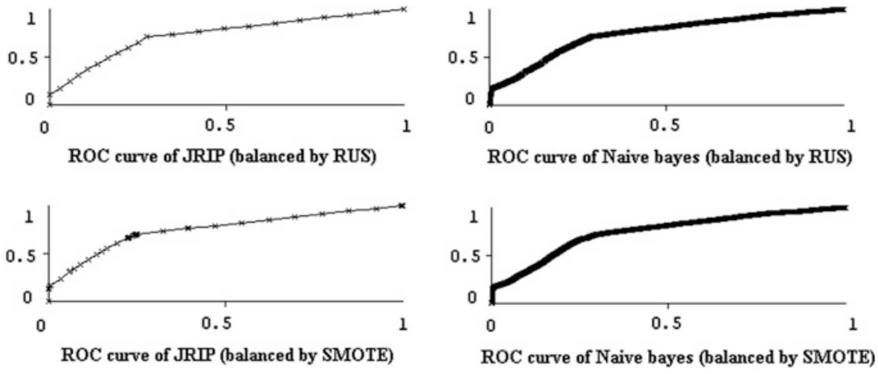


Fig. 2 ROC curve of JRIP and Naïve Bayes after applying RUS & SMOTE

'IBK' (K-nearest Neighbor Classifier) is a classifier that uses the theory of distance based weighting. It tries to find the nearest neighbors depending on the cross validation. It depends solely on how many nearest neighbors are considered during calculation. 'SMO' (Sequential Minimal Optimization) is a 'SVM' (Support Vector Machine) algorithm used for classification. It solves quadratic problems by dividing them into smaller parts and uses analytical approach to solve them. ROC curves for the implementation of IBK and SMO are shown in Fig. 3. As for parameter in IBK, we have used k = 5, seed = 1, and Euclidean distance is used for distance calculation.

'MP' (Multilayer Perceptron) is an artificial neural network that uses a graphical approach to adapt to a selective problem. It is constructed as a graph of nodes with connected edges with weights which are updated during learning. The update occurs during 'back propagation' a method used by the MP to minimize the loss function. 'PART' is classification technique that uses the divide and conquers

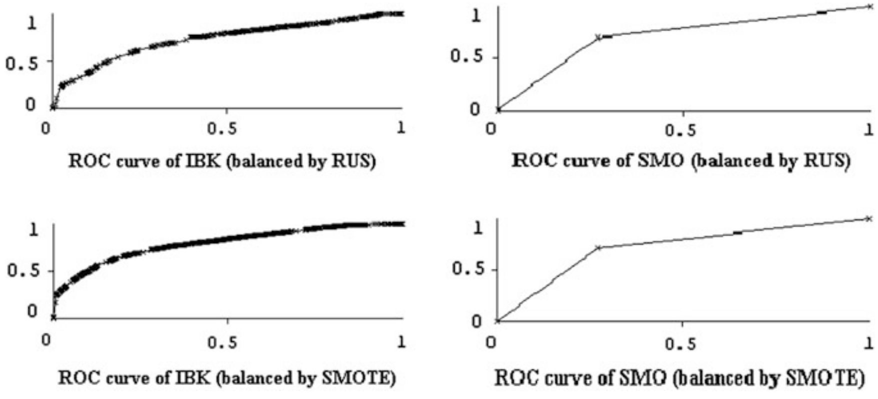
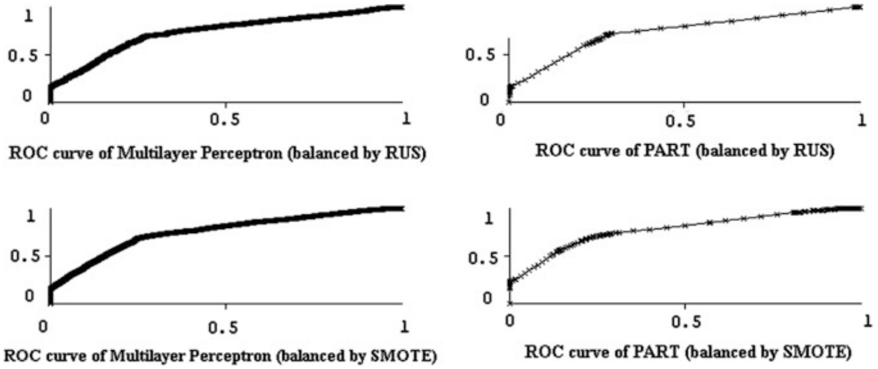


Fig. 3 ROC curve of IBK and SMO after applying RUS & SMOTE



ROC curve of Multilayer Perceptron (balanced by RUS)

ROC curve of PART (balanced by RUS)

ROC curve of Multilayer Perceptron (balanced by SMOTE)

ROC curve of PART (balanced by SMOTE)

Fig. 4 ROC curve of multilayer perceptron and PART after applying RUS & SMOTE

strategy. It generates a partial tree on the present instances and creates rule on the generated tree. ROC curves for the implementation of MP and PART are shown in Fig. 4.

To compare the results we consider FP rate, precision, F-measure, AUC for ROC, correctly classified instances which are well known factors for evaluating a classifier. FP rate considers the false positive rate of a classifier from its confusion matrix. It designates that how many times the classifier predicts a false class as true class. Precision indicates general precision of the classifier which is the ratio of total records and number of correct classifications. F-measure is the harmonic mean of recall and precision with a weighted value. AUC stands for area under the curve. It calculates the area under ROC curve to show the performance of a classifier. Correctly classified instances are just the percentage of correctly classified instances from total classified instances.

5 Evaluation

The training set and testing set which is needed for classification techniques are generated by cross validation technique. We use tenfold cross validation technique as it is better than other techniques to validate the dataset (Figs. 5 and 6).

FP rate and correctly classified instances are two very important factors in determining the performance of a classifier. Lower FP-rate and higher correctly classified instances represent better efficiency of a classifier. From our experiment we see that IBK has the highest FP-rate (0.319) on RUS but is lower in correctly classified instances (68%) so it is one of the poorest performing classifier. RIPPER is efficient in correctly classified instances (73%) and is also lower on FP-rate (0.27) for SMOTE, and for RUS it has almost the same ratio. SMO works moderately on both RUS and SMOTE. Naive Bayes has a high FP-rate on RUS (0.31) and is also less efficient on correctly classified instances (68%). MP has an average value of

Fig. 5 Comparison of the performance of different classifier based on FP Rate

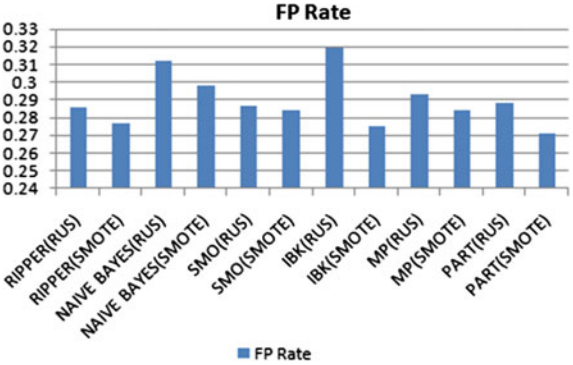
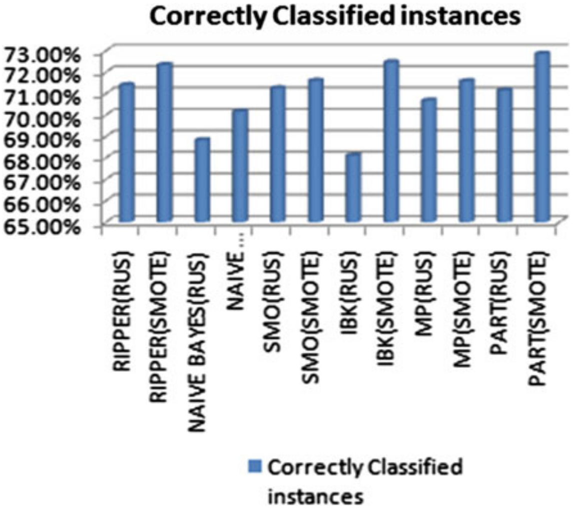


Fig. 6 Comparison of the performance of different classifier based on correctly classified instances



FP-rate and correctly classified instances compared to others. From the results it is evident that balancing the data with SMOTE and using PART classification on the balanced data gives the maximum correctly classified instances (72.86%) and lowest FP-rate (0.271). It is also high on precision (0.732) and on F-measure (0.728) and AUC of ROC (0.773). The performance of PART algorithm is significant in this dataset compared to others.

6 Conclusions

Our goal is to find a classifier that could effectively classify a non-regular customer from a regular customer of an insurance company. To do this initially we face some problems in preprocessing stage. To solve this we use first attribute selection methods to select the proper attributes that can have maximum effect on the classification. It proves to be very effective in action. We also use balancing algorithms on our data to balance the data. Without applying balancing techniques the classification is mostly favored by the general class. But after balancing the results that we get are quite good. As the balancing is done maintaining the initial ratio, the result is equally applicable on original data set.

7 Future Work

Data mining techniques are very useful to apply on life insurance companies data. The regularity of a customer for installment payment depends on certain important factors that the company stores which are obviously user specific and very sensitive. The company that provided us the data could not provide user specific information such as the actual income of the policy holder, health condition of the policy holder etc. which can be integrated in the attributes effecting classification of a customer. We intend to collect these user sensitive information which we believe will effect strongly in building a more specific and effective classifier in future.

References

1. Kirlidog, M., Asuk, C.: A fraud detection approach with data mining in health insurance. Proc.-Soc. Behav. Sci. **62**, 989–994 (2012)
2. Xiaoyun, W., Danyue, L.: Hybrid outlier mining algorithm based evaluation of client moral risk in insurance company. In: 2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME), pp. 585–589. IEEE (2010)
3. Yan, Y., Xie, H.: Research on the application of data mining technology in insurance informatization. In: Ninth International Conference on Hybrid Intelligent Systems, 2009. HIS’09, vol. 3, pp. 202–205 (2009)
4. Goonetilleke, T.O., Caldera, H.A.: Mining life insurance data for customer attrition analysis. J. Ind. Intell. Inf. **1**(1)
5. Thakur, S.S., Sing, J.K.: Mining customer’s data for vehicle insurance prediction system using k-means clustering—an application. Int. J. Comput. Appl. Eng. Sci. **3**(4), 148 (2013)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 321–357 (2002)