Inducing Valuable Rules from Imbalanced Data: The Case of an Iranian Bank Export Loans

Seyed Mahdi Sadatrasoul
Ph.D student of Industrial Engineering, Iran University of Science and Technology (IUST)
IUST, Farjam St., Tehran, Iran
Sadatrasoul@iust.ac.ir

Mohammad Reza Gholamian
Assistant prof. of Industrial Engineering, Iran University of Science and Technology (IUST)
IUST, Farjam St., Tehran, Iran
Gholamian@iust.ac.ir

Kamran shahanaghi
Assistant prof. of Industrial Engineering, Iran University of Science and Technology (IUST)
IUST, Farjam St., Tehran, Iran
Shahanaghi@iust.ac.ir

ABSTRACT

Credit scoring is a classification problem leading to introducing numeroustechniques to deal with itsuch as support vector machines, neural networks and rule-based classifiers. Rule bases are the top priority in credit decision making because of their ability to explicitly distinguish between good and bad applicants.

In a credit- scoring context, imbalanced data sets frequently occur as the number of good loans in a portfolio, which is usually much higher than the number of loans that default.The paper is to explore the suitability of RIPPER, OneR, Decision table, PART and C4.5 for loandefault prediction rule extraction.

A real database of one of Iranian banks export loans is used, and class imbalance issues are investigated in its loan database by random oversampling the minority class of defaulters along with three sampling of majority in non-defaulters class. The performance criterion chosen to measure such an effect isthe area under the receiver operating characteristic curve (AUC), accuracy measure and number of rules. Friedman's statistic is used to test significant differences between techniques and datasets. Theresults shows that PART is the best classifier in all of balanced and imbalanced datasets.

Keywords

Credit scoring, Banking industry, Rule extraction, Imbalanced data sampling

1. INTRODUCTION

In today's competitive economy, credit scoring is widely used in banking industry. Every day, individual's and company's records of past borrowing and repaying actions are gathered and analyzed by information systems. Banks use this information to determine the individual's and company's profit.Application (credit) scoring is one of the main issues in the process of lending [1].In this paper we will address the credit scoring problem.Credit scoring is used to answer one key question: what is the probability of default within a fixed period of year. Credit scoring derives from banks historical loans data to classify customer as good or bad.

There are many techniques suggested to perform classification in the credit scoring problems, includingstatisticaland intelligent techniques. Logistic regression is the most favorite statistical and traditionalmethod used to assess the credit score[2].Lineardiscriminating analysis is also applied and shows that it is as efficient as logistic regression [3]. There are also many intelligent techniques applied to the problemincluding neural networks, Bayesian networks, support vector machines, case-based reasoning, decision trees etc. Some studies have shown that neural networks, SVM, decision-making trees and other intelligent techniques are superior to statistical techniques [4-6].

In recent years hybrid techniques are also proposedsince theyare the main focus of many researchers.Hybrid techniques usually use different algorithms strengths to improve the other algorithms weaknesses. In some hybrid techniques both statistical and intelligent techniques are used.Besides,different hybridization algorithms are used in the literature. A hybrid neural discriminant technique with BP neural network and discriminant analysis are proposed, indicatingmore accuracy than the BP neural network and discriminant analysis[7].A two-stage hybrid procedure with artificial neural networksand multivariate adaptive regression is also proposed[8]. In a study hybrid approaches are divided into four main areas. To achieve the goal, different combination of clustering algorithms and classifiers are tested. Accordingly,logistic regression and neural network hybriddemonstrated the best accuracy[9].In other studies a hybrid Meta heuristic techniques with intelligent techniques is used.An integration of support vector machines, genetic algorithms and F-score is studied[10]. In the last decade, using Ensemble techniques increased in the area and in some cases has led to better accuracy rate[11, 12]. Neural network ensemble strategies includingcross validation, bagging and boosting for financial decision applications are studied and shown better accuracy rate and generalization ability[11]. Ensemble learning is an open issue in recent year'sstudies[13, 14].

Because of robustness and transparency needs and also the auditing process done by regulators on the credit scoring in some countries,Banks cannot use many of mentioned techniques [15].By using rule bases, banks can easily interpret the results and explore the rejecting reasons to the applicant and regulatory auditors. There is actually a little literature in the field of rule-\based credit scoring. Ben-Davide provides a new method for rule pruning and examined his method on the credit scoring data set[16]. Hoffmann et.al introduced a new learning method for fuzzy rule induction based on the evolutionary algorithms[17].Martens et al used the support vector machine for rule induction in the credit scoring problems[18].Malhotraet. Al. used the adaptive neuro fuzzy inference systems(ANFIS) for rule induction and showed that this method works better than discriminant analysis on their own credit scoring dataset, which is gathered from credit unions[19].They used the back propagation method to learn their Rules membership function to fit on the data.Baesens et.al. used and evaluate dthreeneural network rule extractiontechniques includingNeurorule, Trepan, and Nefclassfor rule extraction in three real life data bases (German credit database, Bene1 and Bene2 credit database). They showed Nerorule and Trepan yield better classification accuracy compared to the C4.5 algorithm and the logistic regression. Finally, they visualize the extracted rule sets using decision table[20].

In a credit scoring context, imbalanced data sets frequently occur as the number of good loans in a portfolio, which is usually much higher than the number of loans that default[21]. It is reported that defaults ratio are ten percent of the whole bank's loan portfolio on average[22]. As shown in practical studies,the real credit scoring datasets are imbalanced. There are some but few studies, which investigate imbalanced credit scoring data sets. Huang, Hung and Jiau proposed a strategy of data cleaning for handling imbalanced distribution of credit data to overcome problems of over fitting and the relevance of classifiers[23]. Brown and Muesconducted several experiments based on different classifiers on fiveparts of UCI and non-UCI credit datasets; consequently, they balanced their samples on 70(good)/30(bad)[21]. Their experiments show that random forest and gradient boosting classifiers perform very well in the credit scoring context.

The aim of this paper is to conduct a study of various rule- based techniques based on five instances of an Iranian bank export loans data. In order to extract valuable rules bases the results are compared in terms of area under the receiver operating characteristic curve (AUC), accuracy and number of rules.

The study is divided into four other major parts: section 2 describes the classification techniques used. Section 3 introduces the data, experiments settings, Section 4 discusses their results andthe concluding result is studied in section 5.

2. Overview of classification techniques

The paper aims to extract the best rules from imbalanced data in the credit scoring context. For this purpose 5 rule-based and tree induction (with the aim of rule induction) classifiers are selected. A brief description of these techniques is presented below.

2.1. C4.5

Decision trees split the data into smaller subsets using their nodes and at the end of each node a series of leaf nodes assigning a class to each of the observations. C4.5 built trees based on the concept of information theory[24] in which entropy of a sample of K, can be computed by:

Entropy (k) = $-p_1 log_2(p_1) - p_2 log_2(p_2)$(1)

Where $p_1(p_0)$ are the proportions of the class values 1(0) in the sample K. The attribute with the highest normalized information gain is used for this division. The algorithm is used on the smaller subsets iteratively.

2.2. RIPPER

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is a rule-based learning that builds a setof rules by minimizing the amount of error[25]. In the optimization step if the modified rule is better according to an MDL,heuristic rules are replaced with a modified one in order to reach a small rule set.

2.3. One R

OneR is a one-level decision tree algorithm, which selects attributes one-by-one from a dataset and generates a different set of rules based on error rate. At last, the attribute and its appropriate rule set with minimum error is selected[26].

2.4. Decision table

Decision Table algorithm build tables using a simple decision table majority classifier[27]. Itusesa 'decision table' to summarize the dataset.After finding the line in the decision table that fits the non-class values, a new data item is assigned a category. Thenthe wrapper method isemployed to find a good subset of attributes for inclusion in the table. The likelihood of over-fitting is reduced by eliminating attributes that contribute little or nothing to a model of the dataset and at last a smaller, well-defined decision table is reached.

2.5. PART

Partial decision tree algorithm (PART) is a developed version of RIPPER and C4.5[28]. Its main improvement is that it does not need to perform global optimization like C4.5 and RIPPER to produce rules. It uses the standard covering algorithm to generate a decision list, and avoids over- pruning by inducing rules from partial decision trees.

3. RESEARCH METHODOLOGY

3.1. Data sets characteristics

An Iranian commercial bank real export loan dataset is used to evaluate the proposed algorithm. Table (1) shows the characteristics of the dataset. The initial dataset include 1109 corporate applicants and46 financial and non financial data in the period from 2007 to 2012. First, the data cleaning is done; it includes removing redundant, outliers'data and missing values. There were a few missing values for some corporate: some of them lack financial data and the others lack the result of their loans.In fact, in the process of debt repay, some of them are not applied for loan yet. As a result,387corporate are excluded. From 722 remainedcorporate,652 are credit worthy (90.3%) and other 70 was unworthy (9.9%). Dummy variables were created for the categorical variables (ex. Type of industry).Using dummy variables; thenumber of variables increased to 55.Table (1) summarizes the dataset characteristics before and after cleaning step.

Table 1.dataset description

status	Data size	Inputs variables		
		Total	Continuous	Categorical
Before cleaning	1109	46	38	8
After cleaning	722	55	34	21

Delinquency status was defined by Basel committee definition of "default" and used to generate a 1/0 target variable for modeling purposes (good = 1, bad = 0). Accounts with no more than three months or more in arrears were classified as good. Those that were currently three or more months in arrears, or had been three months in arrears, were classified as bad.The results and descriptions of the variables used are shown in table (5) inappendix (1).

3.2. Re-sampling setup

Table (2) shows the main imbalanced dataset and samples built in order to consider imbalanced issue. The main dataset has a 90/10 class distribution, a 75/25 class distributions selected for balancing the data and the main database is altered in different scenarios to meet this distribution.The two most common preprocessing techniques are random minority oversampling (ROS) and random majority under sampling (RUS). In ROS, instances of the minority class (bad applicants) are randomly duplicated in the dataset. In RUS, instances of the majorityclass (good applicants) are randomly discarded from the dataset.

In this study four different balanced datasets are created using two mentioned techniques. First, using ROS bad instances are duplicated and the "Oversampled dataset" is created. This duplication is done until the distribution of good/bad meets to 75/25 so the number of bad instances increased from 70 to 217 samples.In another re-sampling scenario, using RUS, three different "Under sampled datasets" are created. In order to use all of the datasets, simple random sample without replacement is done. The 'under sampled dataset' are designed in a manner that each good applicant in the main dataset is included in one and only one of three different 'under sampled datasets ' is selected. This reduction is done until the distribution of good/bad meets nearly to 75/25, so the number of good instances reduced to these three under sampled datasets sequentially to 218,226 and 208 samples.

Table 2.Different samples of dataset used

Dataset name	Data size	Good	Bad	Good/All percent
Main imbalanced dataset	722	652	70	90.3
Oversampled dataset	869	652	217	75.02
Under sampled dataset No.1	288	218	70	75.74
Under sampled dataset No.2	297	226	70	76.9
Under sampled dataset No.3	278	208	70	74.82

3.3. Performance analysis

Five different measures are used to analyze the performance of the constructed rule bases. The performance criterion chosen to measure the effect of significant difference in number of observations is the area under the receiver operator characteristic curve(AUC) statistic[21].Confusion matrix is another favorable instrument used in performance evaluations as shown in table (3).Overall accuracy, Good precision and bad precision are important measures after the ROC measure, as they show the classifications quality from other dimension.

Table 3.The confusion matrix

	PREDICTED CLASS		
		Class= Worthy	Class= Unworthy
ACTUAL CLASS	Class=Worthy	a(TP)	b(FN)
	Class= Unworthy	c(FP)	d(TN)

The overall accuracy of successfully identifying loans is computed using equation (2)

$$\text{Overall accuracy}=\frac{TP+TN}{TP+TN+FN+FP} \qquad (2)$$

Theprecision of successfully identifying non-default loans is computed using equation (3)

$$\text{Good precision}=\frac{TP}{TP+FP} \qquad (3)$$

Theprecision of successfully identifying default loans is computed using equation (4)

$$\text{Bad precision} =\frac{TN}{TN+FN} \qquad (4)$$

Compactness of rules is another issue in rule base systems. At a defined level of ROC and accuracy measures for two rule bases, the rule base which has lower number of rules is preferred.

4. RESULTS AND DISCUSSIONS

All the experiments in this paper are done using 10 fold-cross validation. Table (4) shows classification accuracy, number of rules and area under curve for five datasets. The best classification accuracy, the lowest number of rules and area under curve for each data set are bolded. The best results for all of experiments are also underlined. Three groups of experiments are done and their results are presented below:

4.1. Group one experiment

First atest set at the 5% level of importance from the best performer using Friedman’s test is doneagainst different datasets for all of performance measurements as follows:

- It shows that the results of oversampling data set have a significant difference rather than other four datasets; it can be concluded that oversampling and increasing the number of observations has better results than the other reduction techniques at a defined level of good/bad ratio(75/25).
- The three 'under sampled datasets ' haven’t any significant difference in their results; it can be concluded that different good observations in three different datasets don’t have an import issue in the results.
- The main dataset and three 'under sampled datasets ' haven't shown any significant difference; another separated Friedman test for AUC confirms this hypothesis.
- The Number of rules does not showa significant change in all of the datasets and techniques, excluding decision table. It shows a significant difference and an increase in number of rules in oversampled dataset.

4.2. Group two experiments

Table 4.Performance measures on different datasets and classifiers

dataset	Method	AUC	Accuracy(ALL)%	Precision(Bad)%	Precision (good)%	Number of rules
Main imbalanced dataset	RIPPER	0.531	89.47	**31.3**	90.8	2
	Decision table	0.499	**90.3**	0	90.3	**1**
	OneR	0.494	89.20	0	90.2	3
	PART	**0.612**	87.40	27.7	**91.6**	28
	C4.5	0.574	87.11	20.5	90.9	19
Over sampled dataset	RIPPER	0.881	87.45	72.3	93.3	**15**
	Decision table	0.887	80.21	57.5	92.3	575
	OneR	0.643	76.87	55.2	81.5	45
	PART	**0.941**	**90.22**	75.8	**96.2**	22
	C4.5	0.93	90.1	**76.1**	95.8	48
Under sampled dataset No.1	RIPPER	0.594	72.92	37.5	77.3	3
	Decision table	0.492	**73.95**	0	75.3	**1**
	OneR	0.544	73.61	40	**77.5**	7
	PART	**0.667**	72.22	**42.6**	71.4	22
	C4.5	0.595	69.79	36.1	78.9	24
Under sampled dataset No.2	RIPPER	0.517	73.99	34.8	77.3	**1**
	Decision table	0.511	**75.67**	25	76.4	**1**
	OneR	0.518	71.62	29.4	77.1	6
	PART	**0.656**	71.28	**38.8**	**80.8**	17
	C4.5	0.535	69.93	32.7	78.4	25
Under sampled dataset No.3	RIPPER	0.538	71.94	38.9	76.9	2
	Decision table	0.525	**73.02**	22.2	74.7	**1**
	OneR	0.504	71.22	27.3	75	7
	PART	0.581	71.58	**42.4**	**79.5**	20
	C4.5	**0.596**	68.70	38	79.2	20

5. CONCLUSION

In this paper, a number of different classifiers are used and compared on various balanced and imbalanced datasets. The techniques include RIPPER,C4.5, PART, OneR and Decision table. Animbalanced dataset from a major Iranian bank is applied and balanced using several random oversampling and under sampling techniques.Classifiers and datasets are compared using five different performance measures and Friedman’s test. The results of the study show that random oversampling of bad loans yield to better performance measurement for all of the classifiers. It is also found that PART classifier is perform better on imbalanced data rather than other classifiers and that it's the best performer in the entire experiments. On the other hand, techniques like OneR and decision table are the worst classifiers.

Next researches can focus on using other oversampling methods and their effect on the classifiers training. Studying the effect of different sampling methods on feature selection also paves the way for the prospective researches.

Appendix (1)

Variables included in Iran credit dataset and their types are shown in table (3).

Table 5.list of variables in Iran commercial bank credit dataset

Variable	type	Variable	type
Net profit	Continuous	Type of industry: industry and mine (=1, other =0)	Categorical
Activeininternal market	Categorical	Type of industry: agricultural (=1, other =0)	Categorical
number of countries that the company export to	Categorical	Type of industry: oil and petrochemical (=1, other =0)	Categorical
Sales growth	Categorical	Type of industry: infrastructure and service(=1, other =0)	Categorical
Target market risk (from 1 to 5)	Categorical	Type of industry: chemical (=1, other =0)	Categorical
Seasonal factors	Categorical	Year of financial ratio	Continuous
Companyhistory(number of years)	Categorical	Type of book: Taxdeclaration(=1,other=0)	Categorical
Top Mangers history	Categorical	Type of book: Audit Organization (=1,other=0)	Categorical
Type of company: Cooperative (=1, other =0)	Categorical	Type of book: Accreditedauditor (=1,other=0)	Categorical
Type of company: Stock Exchange(LLP) (=1, other =0)	Categorical	Inventorycash	Continuous
Type of company:Generic join stock(PJS) (=1, other =0)	Categorical	Accounts receivable	Continuous
Type of company: Limited and others (=1, other =0)	Categorical	Other Accounts receivable	Continuous
Type of company: Stock Exchange (=1, other =0)	Categorical	Stock	Continuous
ExperiencewithBank(number of years in 5 categories)	Categorical	Currentassets	Continuous
Audit report Reliability	Categorical (binary)	Non-current assets	Continuous
Current periodsales	Continuous	Totalassets	Continuous
Prior periodsales	Continuous	Short-termfinancial liabilities	Continuous
Two-Prior periodsales	Continuous	Currentliabilities	Continuous
Current periodassets	Continuous	Long-termfinancial liabilities	Continuous
Prior periodassets	Continuous	Non-current liabilities	Continuous
Two-Prior periodassets	Continuous	Totalliabilities	Continuous
Current periodshareholder Equity	Continuous	Capital	Continuous
Prior periodshareholder Equity	Continuous	Accumulatedgainsorlosses	Continuous
Two-Prior periodshareholder Equity	Continuous	shareholder Equity	Continuous
checking accounts creditor turn over	Continuous	Sale	Continuous
checking Account WeightedAverage	Continuous	Grossprofit	Continuous
Averageexportsover the pastthree years	Continuous	Financialcosts	Continuous
Last three yearsaverageimports	Continuous	worthy/nonworthy) y)	Categorical (binary)

REFERENCES

[1] Ben-David, A.(2008). Rule effectiveness in rule-based systems: A credit scoring case study. *Expert Systems with Applications*,vol. 34, no 4, p. 2783-2788.

[2] Baesens, B.(2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, vol.49, no.3, p. 312-329.

[3] Brown, I.,& Mues,C.(2011). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications.

[4] Crook, J.N., Edelman, D.B .,& Thomas, L.C.(2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, vol.183, no.3. p. 1447-1465.

[5] Cohen, W.W.(1996). Learning Trees an ules with Set-val ed Features.

[6] Dinh, T.H.T.,& Kleimeier,S.(2007). A credit scoring model for Vietnam's retail banking market. *International Review of Financial Analysis*,vol.16, no.5, p. 471-495.

[7] Finlay, S.(2010). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*.

[8] Frank, E.,& Witten, I.H.(1998). Generating accurate rule sets without global optimization.

[9] Huang, Z.(2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, vol.37, no.4 p. 543-558.

[10] Holte, R.C.(1993). Very simple classification rules perform well on most commonly used datasets. *Machine learning*, vol. 11, no.1, p. 63-90.

[11] Hoffmann, F.(2007). Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European Journal of Operational Research*, vol.177, no.1, p. 540-555.

[12] Huang, C.L., Chen, M.C.,& Wang,C.J.(2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, vol. 33, no.4, p. 847-856.

[13] Harrell, F. E.,& Lee, K. L.(1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. Biostatistics: Statistics in Biomedical; Public Health; and Environmental Sciences. The Bernard G. Greenberg Volume. New York: North-Holland, p. 333–343.

[14] Huang, Y.M., Hung,C.M .,& Jiau, H.C.(2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, vol.7, no.4, p. 720-747.

[15] Kohavi, R.(1995). The power of decision tables. *Machine Learning*: ECML-95, p. 174-189.

[16] Lee, T.S.(2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, vol.23, no. 3 p. 245-254.

[17] Lee, T.S.,& Chen, I.(2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, vol.28, no.4, p. 743-752.

[18] Louzada, F.(2011). Poly-bagging predictors for classification modelling for credit scoring. Expert Systems with Applications: *An International Journal*, vol.38, no.10, p. 12717-12720.

[19] Lee, T.S.(2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, vol.23, no. 3 p. 245-254.

[20] Lee, T.S.,& Chen, I.(2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, vol.28, no.4, p. 743-752.

[21] Martens, D.(2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, vol.183, no.3, p. 1466-1476.

[22] Malhotra, R.,& Malhotra,D.K.(2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research*,vol. 136, no.1, p. 190-211.

[23] Ong, C.S., Huang, J.J.,& Tzeng,G. H.(2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*,vol. 29, no.1, p. 41-47.

[24] Quinlan, J.R.(1993). C4. 5: programs for machine learning. Morgan kaufmann.

[25] Tsai, C. F.,& Chen, M.L.(2010). Credit rating by hybrid machine learning techniques. *Applied soft computing*,vol. 10,no. 2, p. 374-380.

[26] Tsai, C.F.,& Wu, J.W.(2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, vol. 34, no.4, p. 2639-2649.

[27] Thomas, L.C.(2009). Consumer credit models: pricing, profit, and portfolios: Oxford University Press, USA.

[28] Van Gestel, T.,& Baesens. B.(2009). Credit risk management: Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital.Oxford University Press, USA.

[29] Wiginton, J.C.(1980). A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, vol.15, no. 3, p. 757-770.

[30] West, D., Dellana,S.,& Qian,J.(2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research*, vol.32,no.10, p. 2543-2559.