Credit risk assessment with imbalanced data sets using SVMs

Swati Singh

Computer Science Department, Govt. Home Science College, Jabalpur (M.P.), India

---

***Abstract*—*Support Vector Machine* or *SVM* is one of the most popular supervised learning algorithms, which is used for classification as well as regression problems. Support vector machines (SVM) have a limited performance in credit scoring issues due to the imbalanced data sets in which the number of unpaid is lower than paid loans. In this work, we developed an SVM model with more kernels on a set of imbalanced data and suggested two data re sampling alternatives - random over sampling (ROS) and synthetic minority oversampling technique (SMOTE). The aim of this work is to explore the relevance of re-sampling data with the SVM technique for an accurate credit risk prediction rate to the class imbalance constraint. The performance criteria chosen to evaluate the suggested technique were accuracy, sensitivity specificity, error type I, error type II, G-mean and the area under the receiver operating characteristic curve.

Keywords*—*support vector machines, credit risk assessment, random over sampling, imbalanced data sets, SMOTE, performance criteria.

I. Introduction

Credit scoring plays an important role for banking institutions to improve their risk assessment measurement. It has been one of the main fields of application of classification issues and attracted growing attention in recent years (Crook et al., 2007). The construction of a powerful credit scoring model is always an exciting challenge because the possibilities for further improvements are almost endless, especially due to the continuous increase in the complexity of parameters that determine solvency.

The literature has shown that credit scoring is an accurate technique for evaluating and measuring credit risk. Many definitions of credit scoring have been proposed by researchers and fully converge to the same principle, which is the detection of the risk of non-repayment of a loan through the prediction of client default probability (Crook et al., 2007; Hand and Henley, 1997; Thomas et al., 2005). Many modeling alternatives, such as traditional statistical methods and non-parametric methods have been developed to manage credit scoring tasks. Then, more powerful models based on artificial intelligence have become popular among researchers. In fact, in practical credit risk assessment applications, most forecasting models often make wrong decisions because of a lack of default data.

The context of imbalanced dataset classification poses a serious challenge for researchers in credit scoring. The main problem is that the number of insolvent clients is usually much smaller than the number of those who are creditworthy. Therefore, the classifier tends to promote healthy clients in the majority class. In other words, healthy clients could be over-represented in the model and can be identified with high accuracy, but insolvent clients, the minority class, are not properly identified. However, to minimize credit risk, it is more important to identify insolvent clients.

This paper presents a range of experiments on a credit data set that has been artificially modified by using two re-sampling techniques [random over sampling (ROS) and synthetic minority oversampling technique (SMOTE)] and a forecasting method. The method we implemented here is the support vector machines (SVM) with multiple kernels. Results were evaluated according to their matrix of confusion and the area under the receiver operating characteristic (ROC) curve (AUC). Performance measures widely used in credit risk prediction systems are: accuracy, type I error, type II error and AUC (Verikas et al., 2010).

II. LITERATURE REVIEW

A. *Credit Scoring*

Credit scoring is a system aimed at ranking credit applications: Those that have a high probability to meet the financial obligations are classified as ‘good’ and those with a low probability are classified as ‘bad’ (Akkoç, 2012; Lee et al., 2002; West, 2000). The score is defined as a tool for early detection of financial difficulties of borrowers. It draws on a statistical approach and leads to a probabilistic risk analysis. A score is a risk rating, or a default probability. Credit scoring is a technique for determining a linear combination of the following form:

Z $\alpha_1 R_1$ $\alpha_2 R_2$ …….$\alpha_n$$R_n$

With Z: borrower score, R_i: Ratio i of the borrower, α: the weighting coefficient of the ratio R_i.

Thomas et al. (2002) defined credit scoring as a set of decision models and techniques that help lenders in the decision to grant credit. The objective of these models is to assign a score to a potential borrower to estimate the future performance of their loan. It is mainly used by banks to predict the probability of default on individual consumer credits and classify borrowers into default and non-default classes. In addition, Abdou et al. (2008) defined credit scoring as an estimation technique based on quantitative data used by financial institutions to assess the creditworthiness of companies and individuals seeking loans. In other words, this method is essentially a decision-support tool that enables the bank to minimize the risk of insolvency as it provides continuous and rapid monitoring of credit applications. Harris (2013) reported that the main idea of credit scoring involves discriminating between good and bad loan potential borrowers. This is done through quantitative measures of performance and of past loan features to predict the performance of future loans (Thomas et al., 2005).

B. *SVM applications for credit scoring*

SVM are widely used for classification and regression problems because of their promising empirical performance. Recently, many studies have used the SVM in the credit scoring with promising results.

Huang et al. (2007) used three strategies to build a hybrid credit scoring model to assess the applicant's credit score from its input characteristics. They proposed a GA-SVM hybrid strategy combining genetic algorithms (GA) with SVM. This strategy can simultaneously perform the tasks of feature selection and model parameter optimization. Harris (2013) used the SVM algorithm to develop credit scoring models of the Barbados Credit Union. This quantitative approach to credit risk assessment in financial institutions in Barbados is currently an underutilized practice (or non-existent). The credit union conducted this study using the traditional assessment approach when making credit authorization decisions. An analysis of the institution's annual reports indicates that this is a serious situation with regard to doubtful receivables. To settle this situation, Harris (2013) developed a number of scoring models. The results presented in his study suggest that the use of appropriate scoring models is developed and improves the institution's decision-making in loan granting.

Li et al. (2006) developed a credit scoring model using SVM to identify potential applicants for loans. The results showed that the SVM model exceeded the artificial neural networks model in terms of generalization. Huang et al. (2004) studied the performance of the SVM approach in predicting credit scoring. They compared the results generated by SVM with the back propagation neural networks. However, a slight improvement of SVM was observed compared to neural networks.

III. METHODOLOGY

A. *Data*

We opted in this research for studying the management of loans to Tunisian companies in different sectors. The experimental data used for this research are collected from a Tunisian commercial bank. The sample contained 408 companies in different sectors of activity, of which 300 were classified as creditworthy and 108 as non-creditworthy. The dependent variable is a binary variable with two values: 1 to creditworthy borrowers and 0 for non-creditworthy borrowers. The tendency to work with two values was the subject of numerous studies (Lee et al., 2002; Akkoç, 2012; Abdou et al., 2008; Bekhet and Eletter, 2014). The dependent variable 1/0 was used for modelling purposes. Companies that correctly repaid their loans and were never late in payment for 90 days or more were classified as solvent. Those who defaulted for 90 days or more at any time in the life of the loan were classified as insolvent.

Table 1 Variables of the study

	Variables	*Measurement of variables*
	Profitability ratios	
V1	Financial profitability	Net income/net equity
V2	Operating profitability	Gross operating surplus/turnover
V3	Economic profitability	Operating income/economic assets
V4	Net profitability	Net income/sales
	Structure ratios	
V5	Financial autonomy	Shareholders' equity/permanent capital
V6	Structural balance	Permanent capital/fixed assets
V7	WC coverage	Working capital/working capital requirement
V8	Solvency	Net capital/total assets
V9	Asset coverage	Net capital/fixed assets
V10	------------------------------	Long/medium term debt/fixed assets
	Debt ratios	

V11	Financial dependence	Long and medium term debt/permanent equity
V12	Repayment capacity	Long and medium term/cash flow net debt
V13	Debt ratio	Financial charges/turnover
V14	Financial burden	Financial expenses/gross operating surplus
	Ratios of rotation	
V15	Working capital ratio	Turnover/total fixed assets
V16	Inventory turnover ratio	Turnover/net stocks
	Other variables	
V17	Devoted turnover	Movement/sales
V18	Share of funding	Bank commitment/banking system commitment
V19	Study duration of a credit report	Log (study period)
V20	Corporate banking relationship duration	1 if the relationship length ≥ 15 months; 0 otherwise
V21	Guarantees	Log (guarantees)
V22	Size of the company	Log (turnover)
V23	Score: credit line number	Log (score)
V24	Ownership structure	1 if the officer holds more than 50% of the capital; otherwise
V25	Legal form	1= SARL; 0 otherwise

The first step in the development of credit scoring model is to decide upon the input variables. To obtain an accurate credit scoring of quantitative financial data and qualitative data are needed. The following 25 variables (Table 1) are expected to affect the loan repayment capacity of which 03 are binary variables and 22 are quantitative variables:

B. Theoretical background to SVM for credit scoring

We present the SVM essential concepts in the simplest case, to obtain a separating hyperplane. In the case of a linear SVM, the score can be represented as a linear combination of the credit applicant characteristics, (e.g., employment status, marital status, etc.) multiplied by some weight. The equation is written as follows:

$$Z \,\square\, w_1 x_1 \,\square\, w_2 x_2 \,\square \ldots\ldots\ldots \square\square w_n \square x_n \square\square\square b$$

where x is the vector representing the characteristics of the individual, n is the dimension of the input vectors, w and b are model parameters.

Given a training sample in the form of pairs $\{y_i, x_i\}$ with i = 1, ..., n x_i R^2 and y_i {+1, −1}. The x_i are vectors of attributes that may belong to two possible classes: a positive class, denoted +1 and a negative class denoted −1. The y_i therefore represents labels or targets associated with x_i. Then, the variable y takes the value −1 to designate a creditworthy borrower and −1 to designate a non-creditworthy borrower.

The main objective of SVM is to find an optimal hyper plane separator capable of separating data and maximizing the margin between these two classes {+1, −1}, in a classification task.

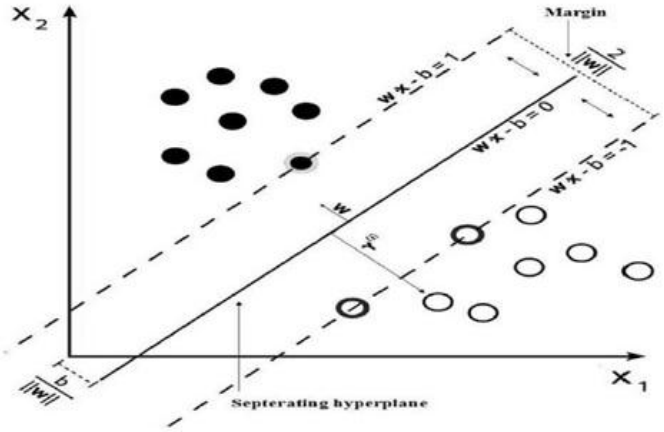

Figure 2. Illustration of support vector machine

The separating hyperplane is represented by the following equation:

$$h(x) \,\square\, w^T x \,\square\, b$$

The hyper plane $w^Tx + b$ represent a hyper plane separating the two classes and the distance between the hyper plane and the closest example is called the margin. Points on the boundaries of the margin are called support vectors. The middle of the margin is called optimal hyper plane separation. The region which lies between the two hyper planes $w^Tx + b = -1$ and $w^Tx + b = +1$ is called the generalization region of the learning machine.

SVM train the parameters w and b from previous training examples of clients that the financial institution had collected over time. This database is composed of a number of examples of clients. Therefore, from a geometrical point of view, the calculation of the values w and b means finding a hyper plane that separates good clients and bad ones. To do so, the SVM maximize the margin between the hyper plane and its closest observations, which are the support vectors from the training data x.

Since this equation satisfies the conditions Karush-Kuhn-Tucker (KKT), the condition $g_i(w) \leq 0$ is an active strain which means it is an inequality.

Therefore, the constraint of the primal problem can be rewritten as:

$$g_i(w) \square y_i \square w^T x_i \square b\square \square 1 - \xi_i \leq 0, i \square 1, ..., n$$

SVMs have several advantages: they have a good capacity for generalization. They can generalize effectively even when they are trained with a few examples. They do not require any background on the possible distribution of the underlying datasets (Chandra et al., 2010). The SVM method requires less effort to designate the appropriate architecture (small number of parameters to be adjusted or estimated). Moreover, it does not require assumptions about the structure of the data such as the normality of the distribution of each of the selected variables and the assumption of independence between them (Lee, 2012). The SVM is based on the principle of finding the separating plane of the largest margin (a hyper plane). This principle minimizes the risk of over-learning of training data and provides a very powerful generalization capability in classification (Bhattacharyya et al., 2011).

C. *The class imbalance problem*

In the case of imbalanced datasets, the prediction methods are dominated by the majority class and, consequently, they poorly classify the observations of the minority class (Vasu and Ravi, 2011). Two techniques were considered in order to restore sampling balance: ROS and SMOTE until the two classes are almost equally represented. The selected sampling strategies are the most popular because they are independent of the underlying classifier and can be easily implemented.

- *Random over sampling*

A way to rebalance the data sets is the random replication of the number of individuals belonging to the minority class as illustrated in Figure 3. The risk of this simplistic approach is to slow the algorithms by adding individuals, while providing models unable to make generalizations (risk of over-training). Previous research (Japkowicz, 2000) discussed over-sampling with replacement and noted that this does not improve the recognition of minority class significantly.

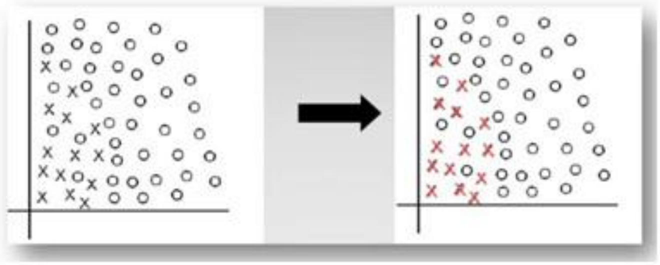

Figure 3 Illustration of the principle ROS (see online version for colours)

- *Experimental approach*

In this paper, the data set is randomly divided into 70% training and 30% test data. The training data are used to build the prediction model and the testing data serves to assess the prediction model efficiency. In fact, the imbalanced training dataset was first used as the input of the SVM algorithm. The second and third procedures concerned ROS and SMOTE. Here, our input dataset was resampled by balancing it in a pre-treatment step. The obtained balanced dataset was used as the input for the SVM algorithm.

- *Performance metrices*

To evaluate different techniques, we used the following performance indicators to compare the results: accuracy (the fraction of correctly classified borrowers), sensibility (the fraction of credit-worthy borrowers correctly classified by the classification model), specificity (the fraction of non-credit-worthy borrowers correctly classified by the classification model), type I error (the fraction of borrowers wrongly classified as insolvent), type II error (the fraction of borrowers wrongly classified as solvent), G-mean (the classification performance balance between majority and minority classes) and area under the ROC curve, i.e., AUC expressed as a percentage of the maximum possible area under an ROC curve. The ROC curve serves to visualise, organise and select classifiers based on their trade-off between false positive rate and true positive rate (sensitivity) (Maalouf and Trafalis, 2011).

$$Accuracy \square \frac{TP \square TN}{TP \square TN \square FP \square FN}$$

$$True\ positive\ rate\ (Sensitivity) \square \frac{TP}{TP \square FN}$$

$$True\ negative\ rate\ (Specificity) \square \frac{TN}{TN \square FP}$$

$$G - mean \square \sqrt{Sensitivity\ Specificity}$$

$$False\ negative\ rate\ (Error\ I) \square \frac{FN}{TP \square FN}$$

$$False\ positive\ rate\ (Error\ II) \square \frac{FP}{TN \square FP}$$

where *TP* (true positive) refers to the number of solvent borrowers correctly classified, *TN* (true negative) refers to the number of insolvent borrowers correctly classified, *FP* (false positive) refers to the number of insolvent borrowers classified as solvent and *FN* (false negative) denotes the number of solvent borrowers classified as insolvent.

IV. Results and discussion

A. Performance comparison

The optimization of the parameters is an important step in the SVM method. Through testing all kernel functions of the original database and the balanced data, the SVM technique appears most efficient in radial kernel as compared to other kernels; accuracy ranking is the most notable. Most often, the credit risk prediction literature applies the RBF kernel because of its versatility, good overall performance and its small number of parameters (C and γ) (Bhattacharyya et al., 2011).

		Accuracy	*Sensitivity*	*Specificity*	*Type I error*	*Type II error*
SVM-RBF	Training	94.410%	0.9952	0.8026	0.0048	0.1974
	Test	82.790%	0.9111	0.5938	0.0889	0.4063
SVM-polynomial	Training	81.120%	0.9905	0.3158	0.0095	0.6842
	Test	80.330%	0.9556	0.375	0.0444	0.625
SVM-linear	Training	84.270%	0.9905	0.4342	0.0095	0.5658
	Test	80.330%	0.9556	0.375	0.0444	0.625
SVM-sigmoid	Training	75.170%	0.9476	0.2105	0.0524	0.7895
	Test	77.870%	0.9556	0.2813	0.0444	0.7188
SVM-RBF	Training	98.81%	0.9857	0.9904	0.0142	0.0095
	Test	92.22%	0.9111	0.9333	0.0889	0.0667
SVM-polynomial	Training	97.86%	0.981	0.9762	0.019	0.0238
	Test	91.11%	0.9	0.9222	0.1	0.0778
SVM-linear	Training	82.62%	0.8524	0.8	0.1476	0.2
	Test	75.00%	0.7889	0.7111	0.2111	0.2889
SVM-sigmoid	Training	71.19%	0.7667	0.6571	0.2333	0.3429
	Test	68.33%	0.9	0.4667	0.1	0.5333
SVM-RBF	Training	99.05%	0.9934	0.9867	0.0066	0.0132
	Test	92.11%	0.9462	0.8878	0.0538	0.1122
SVM-polynomial	Training	84.28%	0.9305	0.7257	0.0695	0.2743
	Test	84.21%	0.9308	0.7245	0.0692	0.2755
SVM-linear	Training	84.28%	0.9172	0.7434	0.0828	0.2566
	Test	82.46%	0.9077	0.7143	0.0923	0.2857
SVM-sigmoid	Training	71.02%	0.8146	0.5708	0.1854	0.4292
	Test	71.49%	0.8538	0.5306	0.1462	0.4694

Comparing the performance indicators of the SVM-RBF method before and after balancing data allowed us to conclude that the rate of correct classification was generally high across all of imbalanced data. This increase is due to the fact that this data set heavily over represented good loans (sensitivity is of the order of 99.52% for the training set and 91.11% for the testing set) and under-represented the number of bad loans (the specificity is in the order 80.26% for the training set and 59.38% for the testing set). This technique shows its inaccuracy to correctly classify minority cases in case of class imbalance. Maalouf and Trafalis (2011) mentioned that accuracy is the most commonly used metric to evaluate the accuracy of the classifier. However, as the results show, accuracy relies more on the majority class and therefore it should not be used as a measure of accuracy for imbalanced data.

The sensitivity value was most similar with the various databases (imbalanced and balanced). However, the results were not significant in terms of the specificity using the original data due to class imbalance (73.53%: 26.47%). In imbalanced classification, the minority class is particularly sensitive to classification errors because of the low number of examples (Moreno-Torres and Herrera, 2010). In the most extreme cases, a single misclassified example of the minority class can lead to a significant decline in performance. So, by introducing the sampling methods (ROS and SMOTE) beside the SVM, we achieved better results in terms of specificity using the ROS method (99.04% for the training set and 93.33% for the testing set) and SMOTE method (98.67% for the training set and 88.78% for the testing set).

The use of resampling techniques can be a good solution to the problem of class imbalance. They constantly improved the forecasting model performance face to imbalanced data. They also had a higher capacity to identify insolvent clients compared to imbalanced data. Also, The SVM method has a high sensitivity to imbalanced data when used for modeling the bank credit risk.

B. Comparison of integrated performances

Our results take into account a single set of credit data because of the confidentiality of the data. However, it would be more beneficial to generalise the experimental results beyond the properties of one empirical dataset. In fact, Crone and Finlay (2012) obtained similar credit scoring datasets across banks and countries. They also claimed that results can be more representative if the sample size and balance are controlled. However, the absence of sufficient additional data sets leads to clear limitations of any empirical experiment.

This work contributed to the improvement of credit scoring systems and therefore increased the deployment of sampling strategies alongside smart techniques that mimic human thought and manage the complexity of real-world problems. Therefore, any improvement in financial forecasting systems can be translated into huge saving (West, 2000). Because of the huge size of the credit markets, even small improvements in classification accuracy could significantly reduce the misclassification costs faced by banks.

CONCLUSIONS

We proposed in this study a new approach using a method of classification and sampling strategies to address the problem of evaluation of client credit at a Tunisian bank, taking into consideration the problems of the class imbalance and improved the accuracy of the identification of insolvent clients.

Although creditworthy clients are identified with great precision, we aimed at finding more effective and efficient methods to solve the problem of evaluation of insolvent credit clients less represented in the sample. In this work, we introduced the SVM forecasting method on two sampling strategies: ROS and SMOTE. The proposed methods were applied to solve a credit risk problem in a bank. We clarified that the proposed approaches can identify insolvent clients more efficiently than the original data. Through the discussion of our study, it was confirmed that the treatment methods of imbalanced data sets with the intelligent technology (SVM) can be applied to solve practical credit problems.

Our results provided an interesting perspective on the role of sampling strategies and artificial intelligence in the Tunisian credit market and their impact on credit risk. In order to improve the process of managing the credit risk of banks in Tunisia, it would be wise to set up templates and sampling strategies beside classification techniques to monitor and control the credit granted.

As a perspective, the cost-sensitive learning is also a potential alternative for modeling and predicting credit risk in a situation of imbalanced data. This needs to be thoroughly investigated because it matches the characteristics of real-world credit risk forecasting.

REFERENCES

[1] Akkoç, S. (2012) 'An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (ANFIS) Model for credit scoring analysis: the case of Turkish credit card data', *European Journal of Operational Research*, Vol. 222, No. 1, pp.168–178.

[2] Bhattacharjee, B., Sridhar, A and Shafi, M. (2017) 'An artificial neural network-based ensemble model for credit risk assessment and deployment as a graphical user interface', *International Journal of Data Mining Modelling and Management*, Vol. 9, No. 2, pp.122–141.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J.C. (2011) 'Data mining for credit card fraud: a comparative study', *Decision Support Systems*, Vol. 50, No. 3, pp.602–613.

[4] Chandra, D.K., Ravi, V. and Ravisankar, P. (2010) 'Support vector machine and wavelet neural network hybrid: application to bankruptcy prediction in banks', *International Journal of Data Mining Modelling and Management*, Vol. 2, No. 1, pp.1–21.

[5] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, Vol. 16, No. 1, pp.321–357.

[6] Crone, S.F and Finlay, S. (2012) 'Instance sampling in credit scoring: an empirical study of sample size and balancing', *International Journal of Forecasting*, Vol. 28, No. 1, pp.224–238.

[7] García, V., Marques., A.I and Sanchez, J.S. (2012) 'Improving risk predictions by preprocessing imbalanced credit data', *19th International Conference on Neural Information Processing (ICONIP 2012), Proceedings*, pp.68–75.

[8] Hand, D.J and Henley, W.E. (1997) 'Statistical classification methods in consumer credit scoring: a review', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 160, No. 3, pp.523–541.

[9] Huang, Z., Chen, H., Hsu, C.J., Chen, W.H. and Wu, S. (2004) 'Credit rating analysis with support vector machines and neural networks: a market comparative study', *Decision Support Systems*, Vol. 37, No. 4, pp.543–558.

[10] Kou, G., Peng, Y and Wang, G. (2014) 'Evaluation of clustering algorithms for financial risk analysis using MCDM methods', *Information Sciences (in press)*, http://dx.doi.org/ 10.1016/j.ins.2014.02.137.

[11] Lee, M.C. (2012) 'Enterprise credit risk evaluation models: a review of current research trends', *International Journal of Computer Applications*, 0975–8887, Vol. 44, No. 11, pp.37–44.

[12] Marqués, A.I., García, V and Sánchez, J.S. (2013) 'On the suitability of resampling techniques for the class imbalance problem in credit scoring', *Journal of the Operational Research Society*, Vol. 64, No. 7, pp.1060–1070.

[13] Moreno-Torres, J.G. and Herrera, F. (2010) 'A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming-based feature extraction', in *Proceedings of the 10th International Conference on Intelligent*

Systems Design and Applications (ISDA'10), Cairo, Egypt, pp.501–506.

[14] Thomas, L.C., Oliver, R.W and Hand, D.J. (2005) 'A survey of the issues in consumer credit modelling research', *Journal of the Operational Research Society*, Vol. 56, No. 9, pp.1006–1015.

[15] Vasu, M and Ravi, V. (2011) 'A hybrid under-sampling approach for mining unbalanced datasets: applications to banking and insurance', *International Journal of Data Mining Modelling and Management*, Vol. 3, No. 1, pp.75–105.

[16] Verikas, A., Kalsyte, Z., Bacauskiene, M. and Gelzinis, A. (2010) 'Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey', *Soft Computing*, Vol. 14, No. 9, pp.995–1010.

[17] Zhang, D and Zhou, L. (2004) 'Discovering golden nuggets: data mining in financial application', *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 34, No. 4, pp.513–522.

[18] Zhou, Z.H and Liu, X.Y. (2010) 'On multi-class cost-sensitive learning', *Computational Intelligence*, Vol. 26, pp.232–257.