A semi-deterministic ensemble strategy for imbalanced datasets (SDEID) applied to bankruptcy prediction

R. A. Mathiasi Horta[1,2], B. S. L. Pires de Lima[1] & C. C. H. Borges[3]
[1]*Civil Engineering Program, COPPE,*
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[2]*Department of Finance and Accounting,*
Federal University of Juiz de Fora, Juiz de Fora, Brazil
[3]*National Laboratory for Scientific Computing, Petrópolis, Brazil*

Abstract

In the last decade, there was a rapid growth in the availability and use of credit for Brazilian companies. Until recently, the decision to grant credit was based on human trial to evaluate the risk of insolvency. Increased demand from companies for credit has led to the use of more accurate models for bankruptcy prediction. In recent years much progress has occurred in the process of drawing up a model fostered by increased competition among financial institutions, changes in the economic environment for businesses and advances in computational techniques. This article discusses and presents alternatives for some of the main problems in the preparation of models for bankruptcy prediction with the application of data mining techniques. The first problem approached is the class imbalance that may cause a poor classification performance and it is treated jointly with an ensemble strategy. The other one rely on the selection of the most significant combination of attributes, the financial variables, which have been widely studied in insolvency prediction. Finally, it is presented a case study in a real world data base of Brazilian companies.

Keywords: data mining, bankruptcy prediction, ensemble, attribute selection, data pre-processing.

1 Introduction

The problem of efficient bankruptcy prognosis is of great interest both to scientists and practitioners. Owners, managers, investors, creditors and business

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080201

partners, as well as governmental institutions, have an interest in assessing the financial position of a firm and its propensity for bankruptcy. Banks need to assess and predict the risk of non recovery of the loan before the extent of it. Financial-statement analysis looks at past and actual firm's financial ratios to predict its future situation. In fact, a model is built based on problem representative cases. Then, the developed model's classification abilities, which depend on the model accuracy, try to respond with correct outputs to real cases or/and new situations.

Despite the modelling techniques that have become increasingly accessible and the rise of the amount of studies in this field, there are several specific problems reported in the literature on the development of bankruptcy prediction models (Balcaen et al [1]). Some of the main problems can be categorized in the following topics: (i) the dichotomy of the dependent variable, (ii) the selectivity of the sample, (iii) the non-stationary and instability of the data, (iv) the use of annual accounting information; (v) the selection of the independent variables and (vi) the temporal dimension of the model.

Besides the problems above mentioned, it can be added another one that regards the Brazilian economic environment: the relatively small number of companies with reliable financial statements for the analyses of this study.

Previous studies have shown that machine learning techniques are superior to those traditional (statistical) methods in dealing with bankruptcy prediction and credit scoring problems (Tsai and Wu [14]).

Most of the problems cited above can affect the performance of the prediction models, and this includes studies dealing with machine learning techniques.

Anyone who has performed data mining on a real-world dataset agrees that knowledge discovery is more than pure pattern recognition. It has been estimated that the actual mining of data only makes up 10% of the time required for the complete knowledge discovery process (Pyle [11]). The precedent time-consuming step of pre-processing is of essential importance for data mining (Han and Kamber [6]). It is more than a tedious necessity: The techniques used in the preprocessing step can deeply influence the results of the subsequent step, the actual application of a data mining algorithm.

This paper deals with some tasks of data preprocessing, as the treatment of the dataset with unbalanced classes and attribute selection. It is also employed a general approach associated with sampling known as ensemble based systems which have shown to produce more accurate results compared to those of single classifier systems.

2 Methodology

2.1 Treatment of imbalanced class data

As stated above, in bankruptcy prediction, the dichotomy of the dependent variable can be regarded as a binary classification problem of two pre-defined groups: bankruptcy or non-bankruptcy. Usually, as it is in our study, the bankruptcy cases constitute a very small amount of the database, revealing an

unbalanced dataset. This unbalanced class distribution cause poor performances using standard classification algorithms without any sampling approach.

Several investigators have analyzed the problem of learning from data sets with unbalanced datasets (Chawla et al [3]; West et al [15], Hung et al [7]). Most methods employ resizing of the training sets, which is simple strategy that includes over-sampling the minority class and under-sampling the majority class. A recent approach to the construction of classifiers from imbalanced datasets employed a synthetic minority over-sampling technique (SMOTE) combined with under-sampling the majority class (Chawla et al [4]).

In this study, we apply those methods for treatment of imbalanced classes combined with ensemble methods described in the following.

In the context of unbalanced datasets, the most suitable performance measures are the ROC curve and f-measure.

2.2 Ensemble methods

An ensemble is a set of classifiers whose individual decisions are combined in some way to sort a set of data whose class is unknown. For Martin Sewell [13] the idea of the ensemble learning is to employ multiple learners and combine their predictions. In his paper, he makes an interesting review of ensemble methods in the literature.

The most employed techniques of ensemble in the risk management literature are bagging and boosting, while the strategy of combining the classifiers is made by majority voting (Hung [7]; Ong et al [10]; Chawla et al [3]).

Bagging (Breiman [2]) was the first effective method of ensemble learning. The method uses multiple versions of a training set by sampling with replacement. Each of these datasets is used to train a different model. The outputs of the models are combined by the average (for regression) or voting (for classification) to create a single output.

Boosting algorithm is one of the most powerful and most used ensemble methods. Schapire [12] stated that a weak classifier (its accuracy on the training set is only slightly better than a random guessing) can be well improved and turned into a strong one through boosting. It can be achieved by an ensemble of models built by re-sampling the data, which are then combined by majority voting.

In majority voting, the outputs of the several numbers of individual classifiers are combined. The output that receives the largest number of votes is selected as the final classification decision (Kittler et al [8]). In general, the final classification decision that reaches the simple majority (greater than half) of votes is taken. In other versions, the ensemble chooses the class on which all classifiers agree, as a unanimous voting or the ensemble chooses the class that receives the highest number of votes, independently if the sum of those votes exceeds 50%.

In this work, it is developed a new methodology of ensemble involving bagging and majority voting as a semi-deterministic ensemble strategy for imbalanced datasets.

2.3 A semi-deterministic ensemble strategy for imbalanced datasets (SDEID)

Inspired on bagging ensemble method, a specific ensemble strategy for prediction with imbalanced datasets is proposed here. It consists of individual classifiers (base classifiers) obtained by different balanced training sets that are generated through re-sampling techniques from the original imbalanced training set.

The size of each new training set is defined according to the number of instances in each class of the original training set. This fact defines a deterministic and a random component in the construction of the balanced training sets.

```
Begin
        Define the number of base classifiers n_bc
        Define the number of instances in each class n_ic
        % construction of the n_bc training sets
        For i=1, n_bc
        % minority class
                Str_i ← Str_m
                %complete by means of a bootstrap process on minority class
            For j = #(Str_m) ,n_ic
                Str_i ← Str_i ∪ bootstrap sample from Str_m
            End
    %majority class
                Str_i ← Str_i ∪ sampled #(Str_M)/n_bc instances from Str_M without
                replacement
                  %complete by means of a bootstrap process on majority class
                  For j = #( Str_M)/n_bc + 1 , n_ic
                  Str_i ← Str_i ∪ bootstrap sample from Str_M
                End
        End
    Training the n_bc base classifiers
    Apply majority voting to classify test data
End.
```

Figure 1. SDEID pseudo-code.

The main idea of SDEID is to include all the examples of the minority class in the new balanced datasets that will train the base classifiers. Since it is interesting to create classifiers whose decision boundaries are adequately different from each other, two approaches are considered to generate diversity of the ensemble process. Initially, a proportional part of the majority class instances, in relation to the number of base classifiers, is randomly distributed without replacement in the balanced training sets. So, it is assured that all majority class instances will be included in the training sets. It is crucial for the strategy to manipulate the classes separately in order to control the class balancing and the instances incorporation. The second step is performed in order to complete, if necessary, the instances related to each class according to the previously determined size of the balanced training set.

An inspired bagging algorithm strategy is used where the necessary instances generated to complete each class in the training sets are obtained randomly with replacement. Thus, the diversity is also assured as bagging algorithm claims. In addition, the deterministic step guarantee that all instances of the original training set participate of the balanced training sets enabling the construction of an ensemble with a low number of base classifiers.

The SDEID pseudo-code is described below. The original training set (*Str*) is composed by the union of the minority class instances (Str_m) with the majority class instances (Str_M), $Str = Str_m \cup Str_M$.

It is assumed that $\#(Str_M) > \#(Str_m)$, where $\#(*)$ is the cardinality of the dataset. As mentioned previously, a minimum value for the number of instances of each class (n_{ic}) in the balanced training sets must be defined:

$$n_{ic} \geq \max(\#(Str_m), \#(Str_M)/n_{bc})$$

where n_{bc} is the number of base classifiers. To high values of n_{bc} SDEID turns closer to bagging algorithms.

It can be observed in the SDEID pseudo-code, elements of bagging, under-sampling and over-sampling methods to build the proposed ensemble algorithm. However, it must be stressed that it is useful to consider the instances of each class separately to facilitate the balancing of the training sets to obtain the base classifiers.

2.4 Techniques of attribute selection

Since, there are a large amount of variables in the finance field; it would be interesting to analyze our case study of bankruptcy prediction combined with some feature selection approaches. Feature selection is the process of identifying and removing as much as possible the irrelevant and redundant information. In this study, some traditional filter feature selection methods are used directly with the imbalanced dataset and with SDEID in order to analyze its performance.

Three techniques were applied in the attribute selection to evaluate its influence on imbalanced datasets: Principal Components Analysis (PCA), CFS (feature selection based on correlation) (Mark and Holmes [9]) and Consistency - Based Evaluation Subset (Liu and Setiono [5]) to have different methodologies in the selection of attributes.

3 Case study

3.1 Problem description

The case study deals with a real world data base of Brazilian companies in the Sao Paulo Stock Exchange (Bovespa). The number of listed companies in Bovespa used in this study is 116 companies, with 27 insolvent companies (bankruptcy) and 89 solvents companies (non-bankruptcy). The analysed data were collected during a total period of ten years (1996-2006). The used number of financial indices (variables) is equal to 17 and they refer to the year of the bankruptcy and the two years before it, which are considered in the juridical

aspect. Therefore, the total number of variables is 17 times 3, totalling 51 variables.

The analyses were performed in WEKA 3.5.6 (Witten and Frank [16]) while the generation of the new datasets was made by SDEID implemented in Matlab 7.1. The method of classification used was Support Vector Machine (SVM) and it was employed cross-validation with 10 k-folds.

Case 1: classification using the original dataset

In this case, the original imbalanced dataset was employed without any treatment. The classification results are presented by a confusion matrix, ROC area and F-measure in Table 1. It can be observed the good ability to classify the class majority, non-bankruptcy (NB), and an inefficient capacity to classify the class minority, bankruptcy (B).

Table 1: Results of the original dataset.

	B	NB	ROC AREA		F MEASURE	
B	3	24	B	NB	B	NB
NB	7	82	0,516	0,516	0,162	0,841

The ROC area was slightly above average (0,516) and F-measure presented a good result only in the NB class (0,841). Regarding these results, it is clear the need for better treatment of the database, where the most important category, the insolvent, has presented a bad performance.

Case 2: classification using re-sampling approaches

In this case, it is presented the individual results of three base classifiers using SDEID. Table 2, shows the results of the classifiers using the datasets SDEID 1, SDEID 2 and SDEID 3 that were generated adopting n_{bc}=3 and n_{ic}=40.

Table 2: Results of single classifiers and ensemble.

		B	NB	F measure	ROC Area
SDEID 1	B	25	2	0.685	0.845
	NB	21	68	0.855	
SDEID 2	B	25	2	0.667	0.834
	NB	23	66	0.841	
SDEID 3	B	24	3	0.750	0.871
	NB	13	76	0.905	
MV	B	27	0	0.740	0.895
	NB	19	70	0.881	

As expected, it can be observed that these classifiers using re-sampling approaches presented better results than the one using the original imbalanced dataset.

The three SDEID were combined by majority voting (MV). The voting strategy combines the class labels of each instance that are available from the

classifier outputs. The ensemble choose the class that receives the highest number of votes. It can be observed that the ensemble classifier presents a better performance than the single classifiers, but the two single classifiers outperformed the ensemble in the NB class.

3.2 Influence of feature selection

In the next study cases three techniques of feature selection are applied to evaluate its influence on the classification accuracy.

Case 3: feature selection using the original dataset

In this case, a feature selection is applied using the original imbalanced dataset. Table 3 presents the results where Imb_Cf means the use of imbalanced dataset with the CFS feature selection method, Imb_Con using the Consistency method and Imb_PCA using PCA analysis.

Table 3: Feature subset in original dataset.

		B	NB	F measure	ROC Area
Imb_Cf	B	4	23	0.235	0.557
	NB	3	86	0.869	
Imb_Con	B	4	23	0.216	0.540
	NB	6	83	0.851	
Imb_PCA	B	2	25	0.118	0.509
	NB	5	84	0.848	

The results show a small improvement in relation to the results of case 1 without the feature selection but as expected, a low performance.

Case 4: feature selection using the previous balanced dataset

In this case the same three techniques of feature selection were applied using the datasets generated with SDEID. The feature selection was performed after the re-sampling in the datasets as it showed better results than performing it before the re-sampling. These analyses found subsets with a number of attributes between 5 and 8. The results did not show a significant improvement in the classification performance nor does any technique have outstood, even though an ensemble analysis was performed in those datasets (MV_fs).

It can be observed that the ensemble of classifiers using the attribute subset selection has presented a slightly improvement in the results compared to the ensemble of case 2. In fact, this improvement occurred in the classification of the solvent companies.

4 Final remarks

In models for insolvency prediction, it is very relevant the cost of predicting the bankruptcy class. As the databases in those fields are usually unbalanced, an

efficient re-sampling method should be employed in order to reduce the cost of error in the minority class, i.e., the bankruptcy class. Another issue that is interesting to investigate is the selection of the subset of relevant attributes.

Table 4: Feature selection in balanced subsets.

		B	NB	F measure	ROC Area
SDEID1_Cf	B	25	2	0.588	0.778
	NB	33	56	0.762	
SDEID2_Cf	B	26	1	0.553	0.751
	NB	41	48	0.696	
SDEID3_Cf	B	26	1	0.542	0.740
	NB	43	46	0.676	
SDEID1_Con	B	23	4	0.533	0.725
	NB	40	49	0.704	
SDEID2_Con	B	26	1	0.542	0.740
	NB	43	46	0.676	
SDEID3_Con	B	24	3	0.533	0.725
	NB	39	50	0.704	
SDEID1_PCA	B	24	3	0.533	0.725
	NB	39	50	0.704	
SDEID2_PCA	B	23	4	0.535	0.724
	NB	36	53	0.726	
SDEID3_PCA	B	24	3	0.527	0.720
	NB	40	49	0.695	
MV_fs	B	27	0	0.761	0.905
	NB	17	72	0.894	

In this work, we applied both approaches in conjunction with ensemble techniques with the purpose of improving the performance of classifying the bankruptcy class in 116 Brazilian companies. The results showed that the proposed ensemble strategy SDEID can greatly improve the classification accuracy and it is more suitable for financial bankruptcy prediction than single classifiers. In future works, we aim to deeply investigate the financial indices and its association with bankruptcy prediction in the Brazilian scenario.

Acknowledgements

The authors are grateful to the Brazilian Research Agencies, CNPq and CAPES, for the financial support for this research.

References

[1] Balcaen, Sofie; Ooghe, Hubert. 35 Years of studies on business failure: on overview of the classical statistical methodologies and their related problems. The British Accounting Review, 38, 2006.

[2] Breiman, Leo. Bagging predictors. Machine Learning, 24(2), 497–501, 1996.
[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16 321–357, 2002.
[4] Chawla, N. V., Japkowicz, N. Kolcz, A. Editorial: Special issue on Learning from imbalanced data sets. 2004.
[5] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection: A Filter Solution," Proc. 13th Int'l Conf. Machine Learning, pp. 319–327, 1996.
[6] Han J, Kamber M., Data mining: concepts and techniques. Second Edition. The Morgan Kaufmann Series in Data Management Systems, 2006.
[7] Hung, Chihli, Chen, Jing-Hong and Wermter, Stefan, Hybrid Probability-Based Ensembles for Bankruptcy Prediction, Proc. of International Conference on Business and Information, July 11–13, 2007.
[8] J. Kittler, M. Hatef, R.P.W. Duin, and J. Mates, "On combining classifiers, "IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226–239, 1998.
[9] Mark A. Hall and Geoffrey Holmes, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, 2003.
[10] Ong, C.S., Huang, J.-J., & Tzeng, G.-H. Building credit scoring models using genetic programming. Expert Systems with Applications, 29, 41–47. 2005.
[11] Pyle D., Data Preparation for Data Mining, Morgan Kaufmann, San Francisco, 1999.
[12] Schapire, Robert E. The strength of weak learnability. Machine Learning, 5(2), 197–227. 1990.
[13] Sewell., Martin. Ensemble methods. 2007, www.machinelearning.net/ ensembles/methods.pdf [access em 30 out 2007].
[14] Tsai, C. F., Wu J. W. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with applications. 2008.
[15] West, David; Dellan, Scott and Qian, Jingxia. Neural network ensemble strategies for financial decision applications. Computers & operations research 32, 2005.
[16] Witten, I.H., Frank. Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems, 2ª ed. 2005.