Improving Risk Predictions by Preprocessing Imbalanced Credit Data

Vicente García[1], Ana I. Marqués[2], and J. Salvador Sánchez[1]

[1] Dep. Computer Languages and Systems – Institute of New Imaging Technologies
[2] Dep. Business Administration and Marketing
Universitat Jaume I
Av. Vicent Sos Baynat s/n, 12071 Castelló de la Plana (Spain)

Abstract. Imbalanced credit data sets refer to databases in which the class of defaulters is heavily under-represented in comparison to the class of non-defaulters. This is a very common situation in real-life credit scoring applications, but it has still received little attention. This paper investigates whether data resampling can be used to improve the performance of learners built from imbalanced credit data sets, and whether the effectiveness of resampling is related to the type of classifier. Experimental results demonstrate that learning with the resampled sets consistently outperforms the use of the original imbalanced credit data, independently of the classifier used.

Key-Words: Credit scoring; Class imbalance; Classification; Resampling; Finance

1 Introduction

Credit scoring constitutes a major instrument for financial institutions to evaluate credit risk, improve cash flow, reduce possible risks and make managerial decisions [16]. In practice, credit scoring refers to a classification problem where a new credit applicant must be categorized into one of the predefined classes (typically, "good" and "bad" applicants, depending on how likely they are to default with their repayments) based on a number of observed variables or attributes that describe socio-demographic characteristics and economic conditions of the applicant.

The most classical approaches to credit scoring are based on parametric statistical models (e.g., linear regression, discriminant analysis, logistic regression and multivariate adaptive regression splines). However, modern credit scoring has been addressed to implement non-parametric methods and artificial intelligence techniques (decision trees, linear programming, artificial neural networks, support vector machines, evolutionary algorithms, rule learners, etc.). In contrast with parametric statistical methods, these alternative models do not assume any specific prior knowledge, but automatically extract knowledge from training observations.

From the many comparative studies carried out, it is not possible to claim the superiority of a method over other competing algorithms regardless of data characteristics (noise, missing values, skewed class distribution, attribute relevance, etc.), which may significantly affect the success of most classification techniques. Whilst some data

complexities have been widely studied, the low-default portfolio problem (also known as the class imbalance problem) has received relatively little attention so far. Nevertheless, imbalanced class distribution naturally happens in credit scoring where in general, the class of creditworthy applicants vastly outnumbers the class of defaulters [11, 14]. For example, it is common to find that defaulters constitute less than 10% of the whole database. This phenomenon of class imbalance may have most influence on the performance of conventional classification techniques because they assume a relatively well-balanced class distribution and equal misclassification costs [9].

During the last decade, the low-default portfolio problem has attracted growing attention, both to detect fraudulent financial activities and to predict creditworthiness of new credit applicants. In the credit scoring domain, research has mainly focused on analyzing the behavior of conventional prediction models, showing that the performance on the minority class drops down significantly as the imbalance ratio increases [2, 10]. However, only a few works have been addressed to design solutions for imbalanced credit data sets. For example, Vinciotti and Hand [18] introduced a modification to logistic regression by taking into account the misclassification costs when the probability estimates are made. Huang et al. [8] proposed two strategies for classification and cleaning of skewed credit data. One method involves randomly selecting instances to balance the proportion of examples in each class, whereas the second consists of combining the ID3 decision tree with a filter.

Yao [21] carried out a systematic comparative study on three weighted classifiers: C4.5 decision tree, support vector machine and rough sets. The experiments over two credit data sets showed that the weighted models outperform the standard classifiers in terms of type-I error. Within the PAKDD'2009 data mining competition, Xie et al. [20] proposed an ensemble of logistic regression and AdaBoost with the aim of optimizing the area under the ROC curve (AUC) for a highly imbalanced credit data set. Florez-Lopez [6] employed several cooperative strategies (simple and weighted voting) based on statistical models and artificial intelligence techniques in combination with bootstrapping to handle the low-default portfolio problem. Kennedy et al. [10] explored the suitability and performance of one-class classifiers for several imbalanced credit scoring problems with varying levels of imbalance. The experimental results suggest that the one-class classifiers perform especially well when the minority class constitutes 2% or less of the data, whereas the two-class classifiers are preferred when the minority class represents at least 15% of the data. Tian et al. [17] proposed a new method based on the support vector domain description model, showing that this can be effective in ranking and classifying imbalanced credit data.

An exhaustive comparative study of various classification techniques when applied to skewed credit data sets was carried out by Brown and Mues [2]. They progressively increased the levels of class imbalance in each of five real-world data sets by randomly under-sampling the minority class of defaulters, so as to identify to what extent the predictive power of each technique was adversely affected. The results show that traditional models, such as logistic regression and linear discriminant analysis, are fairly robust to imbalanced class sizes.

This paper presents a comprehensive suite of experiments over real-life credit data sets, which have artificially been modified to derive different imbalance ratios (propor-

tion of defaulters and non-defaulters examples), using eight resampling methods and four prediction models. All techniques are evaluated in terms of the AUC, and then compared for statistical differences using the Friedman's statistic and the Bonferroni-Dunn post hoc test. The aim of this study is to explore the suitability of data resampling for accurate prediction of credit risk under the class imbalance problem.

2 Resampling Strategies for Handling Imbalanced Data Sets

Much work has been done to deal with the class imbalance problem, at both data and algorithmic levels. Conclusions about what is the best solution are divergent, but the data level methods are the most investigated because they are independent of the underlying classifier and can be easily implemented for any problem. The most popular strategies at the data level consist of resampling the data to obtain an altered class distribution. This can be done by either over-sampling the minority (positive) class or under-sampling the majority (negative) class until both classes are approximately equally represented. This section provides a brief overview of the resampling methods considered in this work.

2.1 Over-sampling

The simplest strategy to expand the minority class corresponds to random over-sampling (ROS), which is a non-heuristic method that balances the class distribution through the random replication of positive examples. Nevertheless, this method may increase the likelihood of overfitting since it makes exact copies of the minority class instances.

In order to avoid overfitting, Chawla et al. [4] proposed a technique, called Synthetic Minority Over-sampling TEchnique (SMOTE), to up-size the minority class. This algorithm generates artificial positive examples by interpolating existing instances that lie close together. It first finds the k nearest neighbors of the minority class for each positive example; the synthetic examples are then generated in the direction of some or all of those neighbors, depending on the amount of over-sampling required (in the experiments here reported, k is set to 5).

Although SMOTE has proved to be an effective tool for handling the class imbalance, it may overgeneralize the minority class as it does not take care of the distribution of majority class neighbors. As a result, it may increase the overlapping between classes. Numerous modifications to the original SMOTE have been proposed with the aim of determining the region in which the positive examples should be generated. For example, the Safe-Level SMOTE (SL-SMOTE) algorithm [3] calculates a "safe level" coefficient (sl) for each positive example, which is defined as the number of positive cases in its k neighbors. If $sl \approx 0$, such an example is considered as noise; if $sl \approx k$, then the example may be located in a safe region of the minority class. The idea is to direct the generation of new synthetic examples close to safe regions.

Batista et al. [1] proposed a method that combines SMOTE and data cleaning, pursuing to reduce the possible overlapping introduced when the synthetic positive examples are generated. In order to create well-defined classes, after over-sampling the minority class by means of SMOTE, Wilson's editing [19] is applied to remove any example (either positive or negative) that is misclassified by its three nearest neighbors. This method is here called SMOTE+WE.

2.2 Under-sampling

Random under-sampling (RUS) aims at balancing the data set through the random removal of negative observations. Despite its simplicity, it has empirically been shown to be one of the most effective resampling methods. However, this technique may discard data potentially important for the classification process. Consequently, other methods have been designed to provide a more intelligent selection strategy. For example, Kubat and Matwin [12] proposed the One-Sided Selection technique (OSS), which selectively removes only those negative examples that are redundant or "noisy" (majority class examples that border the minority class). The border examples are detected by applying Tomek links, and the redundant cases (those that are distant from the decision boundary) are discovered by means of Hart's condensing [7].

Laurikkala [13] introduced a new algorithm called Neighborhood CLeaning rule (NCL) that operates in a similar fashion as OSS. In this case, Wilson's editing is used to remove negative examples whose class label differs from the class of at least two of its three nearest neighbors. Besides, if a positive instance is misclassified by its three nearest neighbors, then the algorithm also eliminates the neighbors that belong to the majority class. A quite different alternative corresponds to under-Sampling Based on Clustering (SBC) [22], which rests on the idea that there may exist different clusters in a given data set, and each cluster may have distinct characteristics depending on the ratio of the number of positive examples to the number of negative examples in the cluster. Thus the SBC algorithm first gathers all examples in the data set into some clusters, and then determines the number of negative cases that will be randomly picked up. Finally, it combines the selected majority class instances and all the minority class examples to obtain a resampled data set.

3 Experiments and Databases

The aim of these experiments is to evaluate the performance of the resampling algorithms described in Section 2 (RUS, OSS, NCL, SBC, ROS, SMOTE, SL-SMOTE, SMOTE+WE) in the context of credit scoring, and also investigate to what extent the behavior of each technique is more appropriate for each type of learner. The classification methods correspond to four models widely applied to credit risk prediction: nearest neighbor (1-NN) rule, multi-layer perceptron (MLP) and radial basis function (RBF) neural networks, and support vector machine (SVM) with a linear kernel.

Five real-world credit data sets have been taken to test the performance of the resampling strategies and classifiers. The widely-used Australian, German and Japanese data sets are from the UCI Machine Learning Database Repository (`http://archive.ics.uci.edu/ml/`). The UCSD data set corresponds to a reduced version of a database used in the 2007 Data Mining Contest organized by the University of California San Diego and Fair Isaac Corporation. The Iranian data set [15] comes from a modification to a corporate client database of a small private bank in Iran. Each original set, except the Iranian because of its extremely high imbalance ratio ($iRatio = 1{:}19$), has been altered by randomly under-sampling the minority class of defaulters, thus producing six data sets with varying imbalance ratios, $iRatio = \{1{:}4, 1{:}6, 1{:}8, 1{:}10, 1{:}12, 1{:}14\}$. Therefore, we have obtained a total of 25 data sets (see Table 1).

Table 1. Some characteristics of the data sets. The last column contains, for each database, the number of defaulters for each imbalance ratio ($iRatio = \{1{:}4, 1{:}6, 1{:}8, 1{:}10, 1{:}12, 1{:}14\}$)

Data set	#Attributes	#Good	#Bad
Australian	14	307	77 51 38 31 26 22
German	24	700	175 117 88 70 58 50
Japanese	15	296	74 49 37 30 25 21
UCSD	38	1836	459 306 230 184 153 131
Iranian	27	950	50

Ten different runs of five-fold cross-validation have been executed. For each iteration of cross-validation, the training set has consisted of four folds, and the remaining fold has been used as a test set. Each resampling technique (and also no resampling) has been applied to the training data, then the four learners have been constructed from the transformed data set, and each of the classifiers has been evaluated on the test set using the AUC measure averaged across the 50 runs. Statistical significance of differences between the resampling algorithms have been assessed using Friedman's statistic followed by pairwise comparisons with the Bonferroni-Dunn post hoc test [5] at significance levels of 5% and 10%.

In total, 10 five-fold cross-validation trials and 25 data sets give 1250 different training sets. Eight resampling techniques, plus no resampling, have been applied to each of the 1250 training data sets, resulting in $9 \times 1250 = 11250$ transformed data sets, each of which has been used for learner construction. Since there are 4 learners, a total of $4 \times 11250 = 45000$ classifiers have been constructed and evaluated in the experiments.

4 Results and Discussion

Table 2 reports the average AUC values and the average Friedman's ranks of each resampling method for the four classifiers. The results achieved with the imbalanced data sets (IDS) are also included for comparison purposes. For each classification model, the algorithm with the lowest (best) average rank across the five credit data sets is underlined, whereas the one with the highest (worst) rank is highlighted in italics type. As can be seen, the use of the imabalanced data sets yields the poorest results in terms of the average Friedman's rank, independently of the classifier. When comparing the different resampling methods, it is worth noting that there is not a unique algorithm that gives the best results for all classifiers: the NCL method appears to perform better than the remaining schemes when using the 1-NN and MLP models, whereas random oversampling is the best method for RBF classification and the SMOTE+WE algorithm is superior to the others when using a SVM. However, it seems that over-sampling generally behaves better than under-sampling, especially in the case of using some intelligent selection strategy.

After applying the Friedman test in order to discover whether there exist significant differences in the AUC results, the Bonferroni-Dunn post hoc test has been employed

Table 2. Average AUC values and ranks calculated by Friedman's test

	1-NN		MLP		RBF		SVM	
Algorithms	AUC	AVR	AUC	AVR	AUC	AVR	AUC	AVR
IDS	0.6596	*7.22*	0.6722	*7.72*	0.6168	*7.66*	0.5122	*7.26*
RUS	0.7205	3.68	0.7539	3.18	0.7389	3.48	0.5883	3.92
OSS	0.6874	5.60	0.7132	5.76	0.7021	5.60	0.5566	5.20
NCL	0.7240	2.96	0.7541	2.98	0.6959	4.58	0.5382	6.42
SBC	0.6097	6.96	0.6235	6.48	0.6265	6.48	0.5222	6.80
ROS	0.6596	*7.22*	0.7115	6.00	0.7454	3.12	0.5633	5.78
SMOTE	0.7080	4.24	0.7372	4.52	0.7236	5.00	0.6138	2.60
SMOTE+WE	0.7207	2.96	0.7469	3.56	0.7288	4.56	0.6240	1.92
SL-SMOTE	0.7107	4.16	0.7259	4.80	0.7280	4.52	0.5730	5.10

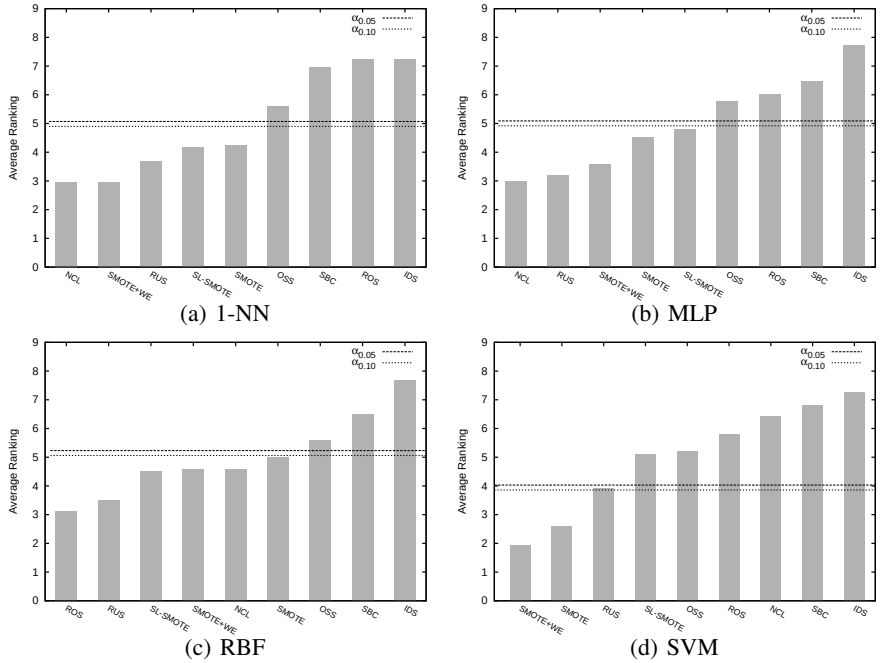
to report any significant differences with respect to the best performing algorithm for each classifier. The results of this test are then depicted to illustrate the differences among the Friedman average ranks. Figure 1 plots the resampling methods against average rankings, whereby all algorithms are sorted according to their ranks. The two horizontal lines, which are at height equal to the sum of the lowest rank and the critical difference computed by the Bonferroni-Dunn test, represent the threshold for the best performing resampling technique at each significance level ($\alpha = 0.05$ and $\alpha = 0.10$). This means that all algorithms above these cut lines perform significantly worse than the best method.

From the Bonferroni-Dunn graphics in Figure 1, one can remark a series of findings. First, in all cases, prediction with the imbalanced data sets is significantly worse than the best performing resampling method, and even worse than the top three algorithms with both $\alpha = 0.05$ and $\alpha = 0.10$. Second, random over-sampling and the OSS and SBC under-sampling methods are generally the worst techniques, showing significant differences when compared with the best algorithms. Third, the performances of random under-sampling, NCL and the SMOTE-based methods are not significantly different.

5 Conclusions

This paper has studied a number of resampling techniques for credit scoring models when addressing the class imbalance problem. The performance of these methods has been assessed by means of the AUC measure, and then the Friedman statistic and the Bonferroni-Dunn post hoc test have been applied to determine whether the differences between the average ranked performances were statistically significant.

The experiments carried out over five real-life credit data sets with varying imbalance ratios have demonstrated that resampling can be a good solution to the class imbalance problem. On the other hand, the results have also allowed to see that NCL under-sampling and SMOTE-based over-sampling outperform the other methods in most cases. The most interesting finding refers to the fact that for all classifiers, the



(a) 1-NN (b) MLP

(c) RBF (d) SVM

Fig. 1. Bonferroni-Dunn graphic for the classifiers

resampling approaches have produced important gains in performance when compared to the use of the imbalanced data sets. In credit scoring applications, a small increase in performance may result in significant future savings and have important commercial implications. Therefore, the improvement in performance achieved by the resampling strategies may become of great importance for banks and financial institutions.

Acknowledgment

Work partially supported by the Spanish Ministry of Education and Science (CSD2007–00018, TIN2009–14205 and AYA2008–05965–0596) and the Generalitat Valenciana (PROMETEO/2010/028).

References

1. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations Newsletter **6**(1) (2004) 20–29
2. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications **39**(3) (2012) 3446–3453
3. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for handling the class imbalanced problem.

In: Proc. of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand (2009) 475–482
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research **16** (2002) 321–357
5. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**(1) (2006) 1–30
6. Florez-Lopez, R.: Credit risk management for low default portfolios. Forecasting defaults through cooperative models and boostrapping strategies. In: Proc. of the 4th European Risk Conference, Nottingham, UK (2010) 1–27
7. Hart, P.E.: The condensed nearest neighbor rule. IEEE Trans. on Information Theory **14**(3) (1968) 505–516
8. Huang, Y.M., Hung, C.M., Jiau, H.C.: Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applications **7**(4) (2006) 720–747
9. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5) (2002) 429–449
10. Kennedy, K., Mac Namee, B., Delany, S.J.: Learning without default: A study of one-class classification and the low-default portfolio problem. In: Proc. of the 20th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland (2010) 174–187
11. Kiefer, N.M.: Default estimation for low-default portfolios. Journal of Empirical Finance **16**(1) (2009) 164–173
12. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Proc. of the 14th International Conference on Machine Learning, Nashville, TN (1997) 179–186
13. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. In: Proc. of the 8th Conference on Artificial Intelligence in Medicine in Europe, Cascais, Portugal (2001) 63–66
14. Phua, C., Alahakoon, D., Lee, V.: Minority report in fraud detection: classification of skewed data. SIGKDD Explorations Newsletter **6**(1) (2004) 50–59
15. Sabzevari, H., Soleymani, M., Noorbakhsh, E.: A comparison between statistical and data mining methods for credit scoring in case of limited available data. In: Proc. of the 3rd CRC Credit Scoring Conference, Edinburgh, UK (2007)
16. Thomas, L.C., Edelman, D.B., Crook, J.N.: Credit Scoring and Its Applications. SIAM, Philadelphia, PA (2002)
17. Tian, B., Nan, L., Zheng, Q., Yang, L.: Customer credit scoring method based on the SVDD classification model with imbalanced dataset. In: Proc. of the International Conference on E-business Technology and Strategy, Ottawa, Canada (2010) 46–60
18. Vinciotti, V., Hand, D.J.: Scorecard construction with unbalanced class sizes. Journal of the Iranian Statistical Society **2**(2) (2003) 189–205
19. Wilson, D.L.: Asymptotic properties of nearest neighbour rules using edited data. IEEE Trans. on Systems, Man and Cybernetics **2** (1972) 408–421
20. Xie, H., Han, S., Shu, X., Yang, X., Qu, X., Zheng, S.: Solving credit scoring problem with ensemble learning: A case study. In: Proc. of the 2nd International Symposium on Knowledge Acquisition and Modeling, Wuhan, China (2009) 51–54
21. Yao, P.: Comparative study on class imbalance learning for credit scoring. In: Proc. of the 9th International Conference on Hybrid Intelligent Systems, Shenyang, China (2009) 105–107
22. Yen, S.J., Lee, Y.S.: Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In: Intelligent Control and Automation. Volume 344 of Lecture Notes in Control and Information Sciences. Springer (2006) 731–740