Ensemble Classifier for Solving Credit Scoring Problems

Maciej Zięba, Jerzy Świątek

Ensemble Classifier for Solving Credit Scoring Problems

Maciej Zięba[1] and Jerzy Świątek[1]

[1] Wroclaw University of Technology, Faculty of Computer Science and Management,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{Maciej.Zieba, Jerzy.Swiatek}@pwr.wroc.pl

Abstract. The goal of this paper is to propose an ensemble classification method for the credit assignment problem. The idea of the proposed method is based on switching class labels techniques. An application of such techniques allows solving two typical data mining problems: a predicament of imbalanced dataset, and an issue of asymmetric cost matrix. The performance of the proposed solution is evaluated on German Credits dataset.

Keywords: credit scoring, ensemble classifier, imbalanced data, cost-sensitive learning

1 Introduction

The insecure financial condition of many institutions in UE and in the USA caused the growing popularity of decision making solutions in bank and financial sectors. Especially accurate decisions about credit assignment are essential for the banks to prevent them from the poor economic condition. Usually, experts from the financial segments are responsible for making credit assignment decisions what generates high costs of maintaining customers. The process of assigning credit status can be automated using methods and algorithms from data mining field. The decision models and their underlying techniques that aid lenders in the granting of consumer credit are known in literature as credit scoring solutions [4].

The key question for decision making about credit status assignment is what characteristics of the consumer should be taken under consideration. According to pragmatism and empiricism of credit scoring the characteristic of the customer (so the vector of the features) should contain only those features, which have meaningful impact on credit decision. Detailed discussion about credit consumer characteristics considered in credit scoring is described in [4].

Another very important aspect of credit scoring (and many other domains, where data mining techniques are applied) is character and quality of the data, which is used to construct decision models. In this work we concentrate on two problems with connected with data: (i) imbalanced data and (ii) asymmetric cost matrix [7]. The problem of imbalanced data is related with disproportions in number of examples from different decision variants (decision classes) in the training data. If we consider the decision problem with two possible decision variants, the imbalanced data problem occurs when the cardinality of examples labeled by one class (called

majority class) is significantly higher than cardinality of examples labeled by the second class (called minority class). The problem of imbalanced data is often considered in parallel to asymmetric cost matrix problem. Such problem can be observed when the cost of classifying object from minority class as an object from majority class is significantly higher than the cost of classifying object from majority class as an object from minority class.

The aim of this work is to propose the decision making algorithm for credit scoring problem, which solves two of the mentioned data mining problems. The problem of making decision about credit assignment is classification task [1] in which the characteristic of the credit consumer is represented by vector of features (also called attributes) $\mathbf{x}_n$ and the set of decision variants is represented by the set class labels $\{C_1, \ldots, C_J\}$. The classification process refers to an algorithmic procedure for assigning a given input into one of a given classes. The algorithm that implements classification is known as classifier, which is denoted by Ψ. The Ψ is build in training procedure, using training set $S_N = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$.

In this work we recommend to use the ensemble classifier [10] that use switching class labels techniques to increase diversity between base classifiers of the ensemble. Our approach is inspired by Breiman's switching class labels technique [3], which was further extended by authors of [12]. In our approach switching probabilities are estimated basing on error rates between classes. According to the proposed procedure it is more probable to switch labels between classes, which are difficult to separate using single classifier and less probable if the classes are almost perfectly separable. Comparing to solution presented in [3] and extended in [12] our approach does not require setting any parameters and maintaining class distribution. In our work we would like to show that switching class labels techniques can be successively applied to deal with problems of imbalanced data and cost-sensitive learning in credit scoring field. Our solution is a alternative to existing solutions, which are mainly based on undersampling and oversampling techniques.

2 Contribution to Value Creation

Nowadays, the crisis on financial markets is observed so it is extremely important for banks and credit institutions to increase their quality rates. The good-quality data mining solutions may help such institutions to make accurate credit assignment decisions which help to reduce the number of dangerous debtors and keep financial status of such companies on the high level.

The proposed classification method is also implemented as a component of Service Oriented Data Mining Systems (SODMS), which is the web data mining system created basing on Service Oriented Architecture (SOA) paradigm. SODMS delivers classification, regression and clustering functionalities as web services [17]. Thanks to universal interfaces the proposed method can be easily used by various types of bank systems without the need of rebuilding the whole system. Such solution reduces

the costs related with software development and makes the bank institution more competitive on the financial market.

3 Related Work

The first scientist, who discovered that the problem of separation "good" and "bad" credits is the problem of finding discriminant function was Durman in 1941 [4]. The growing interest of credit scoring solutions was observed when the credit cards occurred in 1960s but the computational resources were not sufficient to use more sophisticated solutions to deal with the problem. At the beginning of 1990s various data mining techniques were used to estimate the risk of credit approval, especially those, which collects the knowledge in visible form like decision rules and trees [13]. At the beginning of XXI century a growing popularity of ensemble approaches for making credit decisions was observed [9,15]. Such models, which were initialized by the Breiman by proposing bagging algorithm and corresponding statistical framework for the theory of ensembles [2], are powerful tools for solving decision problems which are difficult to be solved using traditional approaches. One of the possible ensemble solutions which can be used to solve credit scoring problem is described in [15]. The authors of this work propose least a squares support vector machines (SVM) ensemble classification model, which combines the benefits gained by combining decision models in ensemble structure with high accuracy of decisions made using SVM. Other ensemble approach for the credit scoring problem is described in [9]. Authors propose to use clustering solutions in preprocessing stage to solve the problem of unrepresentative samples and then they use the ensemble composed of various classification methods to find the final decision about credit assignment. Both of proposed solutions do not touch the problem of imbalanced data and asymmetric cost matrix.

The problem of imbalanced data and corresponding problem of asymmetric cost matrix can be solved by applying oversampling and undersampling techniques [7]. In the simplest case the initial imbalanced dataset can be balanced randomly, either by random sampling objects from minority class and merging them with initial dataset (random oversampling method), or by random selection of the objects from majority class and eliminating them from this dataset (random undersampling method). The random undersampling procedure can be only applied if the distribution of majority class in the training set will not be changed in undersampling process. To save the distribution in undersampling process the procedure of examples selection must be intelligent. One of the possible solutions is informed undersampling, which removes those examples, which are least needed and select only important elements from majority class. Interesting informed undersampling approach is presented in [11]. Authors of this approach present various techniques for imbalanced data problem, which are based on K-NN algorithm.

On the other hand, synthetic samples can be generated in smart way to balance minority class with majority class. Good example of such type of methods is synthetic minority oversampling technique (SMOTE) presented in [5]. This approach uses K-NN to create artificial examples.

Ensembles are also used for imbalanced data problem [6,8]. One of the ensemble solutions for imbalanced problem is SMOTEBoost algorithm [6]. This method uses SMOTE sampling to generate artificial examples for minority class for each of boosting iterations. In such approach, each of created base classifiers concentrates more on minority class. As a consequence, the final classification decision made by ensemble classifier is more balanced. The other example of ensemble approach for imbalanced problem is DataBoost-IM method [8]. This algorithm also uses boosting approach to generate base classifiers. For each of boosting iterations hard examples are identified in current training set. The hard example, which is also called "seed" by the authors, is difficult-to-learn example. Next, each of identified hard examples is used, as a seed, to generate artificial examples. These artificial examples are added to the current training set and the boosting distribution is modified respecting newly added samples.

4 Ensemble Classifier with Switching Class Labels

The typical structure of ensemble classifier is composed of base classifiers on the first level (denoted in this work by $\Psi_1^{(1)}, \ldots, \Psi_K^{(1)}$), which make autonomic class assignment decisions and one combiner (denoted by $\Psi^{(2)}(\mathbf{x})$) situated on the second level of the ensemble which combines decisions gathered from base classifiers and makes the final decision about class assignment. The base classifiers of the ensemble, which can be represented by any simple classification models e. g. decision tree, or neural network, are constructed using datasets $S_{N_1}, \ldots, S_{N_K}$, which are generated from initial training set S_N. Such operation is made to increase diversity of base classifiers what makes the classifier's decisions more independent. In this work we propose the method of building ensemble algorithm which uses switching class labels techniques to increase diversity of base classifiers. This method is based on changing class labels of the objects stored in $S_{N_1}, \ldots, S_{N_K}$, which were generated using typical for ensembles diversification technique (e. g. bootstrap sampling). The operation of class switching is made according to the estimated probability values $\tilde{p}(i|j)$, which represent the probability, that the object, which is a member of j-th class, will be switched to i-th class. It can be observed, that main problem in switching class labels techniques is to find the estimated probability values $\tilde{p}(i|j)$.

Usually, the switching classes techniques are used to increase the diversity of base classifiers, but in this work we focus on using this group of techniques to solve the problem of imbalanced data in parallel with the problem of asymmetric cost matrix for two-class credit scoring problem. Practically it means that we are interested in finding estimated probability values $\tilde{p}(j_{maj}|j_{min})$ and $\tilde{p}(j_{min}|j_{maj})$, where j_{maj} and j_{min} represent majority (positive credit decision) and minority (negative credit decision) class labels respectively. Moreover, we assume that the unit misclassification cost of classifying the object from minority class (negative credit decision) as an object from majority class (positive credit decision) is significantly higher than misclassification cost in opposite direction. To estimate mentioned probability values we evaluate misclassification tendencies between majority and

minority class. To achieve this, the classifier $\Psi_0^{(1)}$ (of the same model as base classifiers of the ensemble) is trained using complete set of examples S_N. Next, the performance of classifier is tested on the same set S_N. During testing procedure, for each pair of class labels $(i|j)$, the number of examples from i-th class classified as member of j-th class group (denoted by $n_{i,j}$) is calculated. Using calculated values $n_{i,j}$, which creates so called confusion matrix, following probability estimators can be constructed:

$$\tilde{p}(j_{maj}|j_{min}) = 0, \tilde{p}(j_{min}|j_{maj}) = \frac{n_{j_{min},j_{maj}}}{n_{j_{min}}} \quad (\mathbf{1})$$

The $n_{j_{min}}$ value represents the number of examples from j_{min} class situated in initial training set S_N. It can be easily observed that switching classes technique is used only for examples from majority class, $\tilde{p}(j_{maj}|j_{min}) = 0$. Such selection of probability estimator is indicated by the asymmetric misclassification costs and was in detailed discussed in [16]. The formal description of the procedure of creating the base classifiers of the ensemble classifier with switching class labels is listed below:

INPUTS:
Training set: $S_N = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$
Number of base classifiers: K

OUTPUTS:
Base classifiers: $\Psi_1^{(1)}, \ldots, \Psi_K^{(1)}$

PROCEDURE:
1. Build classifier $\Psi_0^{(1)}$ on training set S_N
2. Estimate probability value $\tilde{p}(j_{min}|j_{maj})$ by testing $\Psi_0^{(1)}$ on training set S_N
for k **from** 1 **to** K **do**
  3.1 Generate training set $S_{N_{1,k}}$ from S_N using bootstrap sampling without replacement
  3.2 Set $S_{N_{1,k}}^{(\prime)} = \emptyset$
  for i **from** 1 **to** $N_{1,k}$ **do**
    if $(y_i = j_{maj})$
      3.3.1.1 Generate random value r from [0,1]
      if $(r \leq \tilde{p}(j_{min}|j_{maj}))$
        3.3.1.2.1 Set $y_i = j_{min}$
      end if
    end if
    3.3.2 Add example $(\mathbf{x}_i, y_i)$ to $S_{N_{1,k}}^{(\prime)}$
  end for
  3.4 Build classifier $\Psi_k^{(1)}$ on training set $S_{N_{1,k}}^{(\prime)}$
end for

In the first step of the algorithm, classifier $\Psi_0^{(1)}$ is built on training set S_N. The classifier is not the component of ensemble structure, it is created only to identify

misclassification tendencies and, as a consequence, to estimate value of probability $\tilde{p}(j_{min}|j_{maj})$, what is made in the second step of the procedure. Next, the base classifiers $\Psi_1^{(1)}, \ldots, \Psi_K^{(1)}$ of the ensemble are created in the loop in the following way. First, training set $S_{N_{1,k}}$ is generated by bootstrap sampling without replacement from the initial training set S_N. Bootstrap sampling without replacement is sampling with replacement N examples and eliminating the duplicates. Following the procedure, dataset $S_{N_{1,k}}$ is transformed to dataset $S_{N_{1,k}}^{(\prime)}$ using switching procedure. Each object from training set $S_{N_{1,k}}$, which is member of majority class j_{maj}, is switched to minority class j_{min} with the probability $\tilde{p}(j_{min}|j_{maj})$. The training set gained in such way is used to build base classifier $\Psi_k^{(1)}$.

As a second-level classifier, $\Psi^{(2)}(\mathbf{x})$, we propose voting combiner [10], what means, that new object will be classified to the class, which will be selected by majority of base classifiers.

5 Empirical Studies and Future Works

The goal of empirical studies is to evaluate the performance of ensemble classifier with switching class labels described in previous section. The evaluation is made for exemplary credit scoring dataset. The performance of the presented approach was measured with two indexes: (i) empirical risk value and (ii) false negative (FN) rate. The results gained during testing the ensemble classifier with switching class labels are compared with the results achieved by the base classifiers and ensemble approaches, which are commonly observed in classification domain.

The German Credit dataset, which is available in UCI Repository [14], is used to evaluate performance of the proposed ensemble classifier. The data set consists of a set of loans given to a total of 1000 applicants, consisting of 700 samples of creditworthy applicants and 300 samples where credit should not be extended. For each applicant, 20 variables describe credit history, account balances, loan purpose, loan amount, employment status, and personal information. Despite the fact that German Credit dataset is quite old it is still successively used for testing solutions related with credit scoring field [9]. The authors of [9] find The German credit data set very challenging because it is unbalanced and contains a mixture of continuous and categorical values, which confounds the task of classification learning. Moreover, the description of the German Credit dataset recommends using asymmetric cost matrix with the cost of classifying the customer with ”bad” credit status to ”good” class 5 times greater than misclassification in opposite direction.

The ensemble classifier with switching class labels is implemented using WEKA library. The implementation of the classifier is compatible with paradigms of creating data mining services described in [17]. It means that proposed classification method can be published as a web service as a component of the SODMS. As a model of base classifiers Breiman’s Classification And Regression Tree (CART) was selected.

Table 1. Results of empirical evaluation on German Credit dataset for different types of classifiers

Classifiers	ERI value	FN rate
Ensemble algorithm with switching class labels	**0,281**	**23%**
Bagging	0,386	52%
Boosting	0,407	55%
Decorate	0,436	60%
RIPPER	0,442	58%
C 4.5	0,436	56%
KNN	0,453	60%
MLP	0,408	52%
LR	0,393	52%
NB	0,393	51%

The results of empirical studies made on German Credit dataset are presented in Table 1. The performance of ensemble classifier with switching class labels on mentioned dataset was compared with results achieved by classifiers: rule-based classifier (RIPPER), decision tree (C 4.5), K nearest neighbors (K), multilayer perceptron (MLP), logistic regression (LR), Naive Bayes classifier (NB) and ensemble classifiers: bagging, boosting and DECORATE. Two indexes were used to examine the performance: False Negative (FN) rate and empirical risk index (ERI). FN rate is defined as the number of examples from minority class classified as examples from majority class divided by the total number of examples from minority class. ERI can be interpreted as a weighted error value with weights equal to the misclassification costs. The ERI index value achieved by ensemble classifier with switching class labels was 0.1 lower than result gained by bagging, which performed the best among other tested classifiers. The switching class labels techniques implemented in presented approach significantly decrease the empirical risk value achieved on considered dataset. Similar conclusions arise when FN rate is used as comparison index. The value of FN rate for ensemble classifier with switching class labels was equal 23% and was over two times lower than 51%, which was the best result among the rest of tested algorithms. Practically it means, that 50% – 60% customers, which should not obtain the credit, get good credit status when traditional classification approaches are used to make the decision and only 23%, when credit assignment decision is made using ensemble classifier with switching class labels.

The results gained by ensemble classifier with switching class labels significantly better than results achieved by other tested classifier. However, basing on results from one dataset, we can only presume that the proposed classification method outperformed the others by more than 0.1 with respect to ERI. To evaluate the overall performance it is necessary to collect the representative number of datasets and compare the results using statistical methods. Moreover, the ensemble classifier will be adjusted to solve missing values problem in the future works.

Acknowledgments. The research presented in this work has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Bishop C. M.: Pattern Recognition and Machine Learning. Springer, 2006.
2. Breiman L.: Bagging predictors. Machine Learning, 24(2):123–140, August 1996.
3. Breiman L.: Randomizing Outputs to Increase Prediction Accuracy. Machine Learning, 40, 229-242, 2000.
4. Edelman D. B., Lyn C. T., Crook J. N.: Credit scoring and its applications. Society for Industrial and Applied Mathematics, 2002.
5. Chawla N. V., Bowyer K. W., Hall L. O.: SMOTE: Synthetic Minority Over-sampling TEchnique. Artificial Intelligence, 16, 2002.
6. Chawla N. V., Lazarevic A., Hall L. O., Bowyer K. W.: SMOTEBoost : Improving Prediction. Lecture Notes in Computer Science, 2838:107–119, 2003.
7. Garcia E. A.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9):1263–1284, September 2009.
8. Guo H., Herna L. V.: Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach. ACM SIGKDD Explorations Newsletter, 6(1):30– 39, 2004.
9. Hsieh N. C., Hung L. P.: A data driven ensemble classifier for credit scoring analysis. Expert Systems with Applications, 37(1):534–545, January 2010.
10. Kuncheva L. I.: Combining Pattern Classifiers. A John Wiley & Sons, Inc. Publication, 2004.
11. Mani J., Zhang I.: KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In Proceedings of International Conference on Machine Learning (ICML 2003), Workshop Learning from Imbalanced Data Sets, 2003.
12. Martinez-Munoz G., Suarez A.: Switching class labels to generate classification ensembles. Pattern Recognition, 38, 1483-1494, 2005.
13. Quinlan J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann series in machine learning, 1993.
14. UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html
15. Zhou Z., Lai K. K., Yu L.: Least squares support vector machines ensemble models for credit scoring. Expert Systems with Applications 37:127–133, 2010.
16. Zieba M.: Ensemble Methods for customer classification in service oriented systems. Information systems architecture and technology: service oriented networked systems, 2011.
17. Zieba M., Prusiewicz A.: The proposal of service oriented data mining system for solving real-life classification and regression problems. Technological innovation for sustainability: second IFIP, Costa de Caparica, Portugal, 2011.