Credit scoring using SVM and LDA with SMOTE oversampling

Vishv Patel
University of Windsor

Abstract:

In this paper, the problems related to the credit scoring data like class imbalance and data overlapping has been looked upon with various methods like support vector machines, feature selection, one-hot encoding, SMOTE, and linear discriminant analysis. The purpose of this study in the context of credit scoring is to reduce the number of loan defaulting people and that number is shown as false positives in this research to connect it with the performance of a machine learning model via confusion matrix. But since the number of people who default is less than the number who do not default, the problem of class imbalance in the credit scoring is usual. To minimize the number of false positives, the classification model, support vector machine is used in this study with the combination of linear discriminant analysis. The data has been preprocessed with SMOTE to overcome the data imbalanced and to reduce the effect of overlapping in the final result, LDA has been used in this study. Also, the discussions related to, the importance of the models used in this study will be carried out in this paper.

1.Introduction:

Credit scoring is about discriminating the loan payers into good ones (the ones who pay the loan) and bad ones (the ones who does not). The banks have been trying to minimize their loss by minimizing the number of bad loan payers. Now, using machine learning this process can be improvised. In several studies, classification models like SVM, logistic regression, LDA have been used to get more accurate results. Where SVM is found to be successful in classifying the customers that have the chance of getting default in credit card payment with a larger dataset (Bellotti & Crook, 2009). However, the combined approach of SVM and LDA gives better results according to(Xiong & Cherkassky).

It is difficult to make a machine learning model to learn from imbalanced data. Due to the inadequate information of the minority class provided in the imbalanced data, the classification model fails to understand the underlying pattern. Most machine learning model tries to optimize the overall performance without considering the relative distribution of each class, and that is the reason why most machine learning models fails to get better performance in imbalanced data (Liu, Yu, Huang, & An, 2011). To prevent that problem, the oversampling method-SMOTE has been used in this study. According to (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), SMOTE provides more related minority points for the machine learning model to learn from and that is the major reason for getting better results for imbalanced data.
This paper is focused on combining the linear discriminant analysis process with SVM for better separation of classes and preprocessing the data with SMOTE function for oversampling as a solution for imbalanced data.

The further sections are divided into the paper as follows. Section 2 provides a literature review with providing background of similar research done, section 3 focuses on problem definition and section 4 focuses on methodology. Results and discussion are given in section 5 and the conclusion in section 6.

2.Literature review:

Here, the background related to this study will be provided and for the simplicity and

organization, the whole section is divided into 3 main methods used in this study.

2.1 Support vector machine (SVM)

The SVM classifier determines a hyperplane in the space in an attempt to separate the positive class and negative class by maximizing the distance between the hyperplane and the nearest points from both classes. The nearest points from the positive and negative classes are called support vectors.
In (Liu, Yu, Huang, & An, 2011), the effect of imbalance has been included and according to their observation, the SVM is robust and self-adjusting to the moderate amount of imbalance data, while further increase in imbalance can create bias in the boundary towards minority class.

2.2 Linear discriminant analysis (LDA)

The LDA does dimensionality reduction by projecting the data into a lower dimension such that the ratio of between-class distance to within-class distance is maximized. Thus, increasing the class separability (Ye, 2005).

2.3 Synthetic minority oversampling technique (SMOTE)

SMOTE is an attempt to over-sample the data by generating synthetic data points rather than over-sampling it by replacement (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

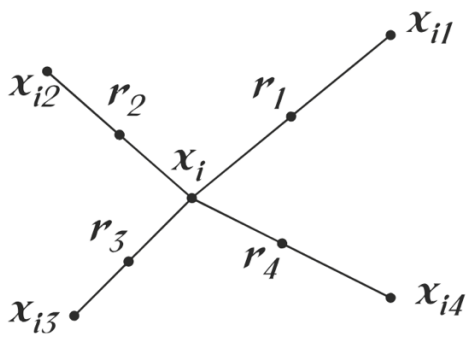

Figure 1: This figure illustrates generating synthetic data points with SMOTE (López, Fernández, García, Palade, & Herrera, 2013).

The xi in fig-1 is the selected data point, while xi1 to xi4 are the nearest neighbors. And r1 to r4 are the synthetic data points.

3. Problem definition:

Here, the original data gives a 68% initial result but with a higher number of false positives. The number of good loan payers in the data is 700 and 300 for the bad loan payers. Even if the machine learning algorithm guesses all the classification to be the good ones, we can get 70% accuracy. But there will also be 30% false positives, which can be bad for this machine learning model as even the slightest number of false positives can cause a financially bad impact on banks. As we discussed in the introduction, most of the machine learning algorithms focus on the overall accuracy without accounting for the relative class distribution in the data (Liu, Yu, Huang, & An, 2011). And, due to a smaller number of minority class data, the machine learning model considers it as noise and fails to include the minority class in generalization (López, Fernández, García, Palade, & Herrera, 2013).

Apart from the problem of data imbalance, the underlying problem of overlapping is also

causing trouble for the machine learning models to getting better results (López, Fernández, García, Palade, & Herrera, 2013). The combination of both imbalanced data and overlapping between classes in the data has a more severe effect on the final results than from any one of them alone (López, Fernández, García, Palade, & Herrera, 2013). It is important to solve the above-stated problems as they are found in many real-life data sources. The need for solving imbalanced learning is growing and this type of learning is found in face recognition, forecasting of ozone levels and medical diagnosis. And also, it is crucial to understand that minority class represents the pattern of interest and it is difficult to extract it from original data (López, Fernández, García, Palade, & Herrera, 2013).
In conclusion, the research for this project is directed from this point towards solving the imbalanced data and overlapping problems to decrease the false positives in the final result.

4. Methodology:

Here, the methodology of this study is being selected to control the imbalance ratio in the data and reduce the overlapping in the data.
The hypothesis for this research is, by solving the problem of imbalanced data and overlapping, the accuracy of the same model can be improvised in terms of reduction in the number of false positives.

4.1 Gathering the data
The data was gathered from the website (Ferreira, 2018). It is an open-source platform for data collection.

4.2 Data:
Here, the data provided has 7 categorical features and 3 numerical features.

4.3 preprocessing

The data with categorical features are encoded by one hot encoder and the missing data is treated by KNN imputer. The feature selection has been performed in the data for removing the features that do not give a contribution to the final result. Than to treat the data imbalance problem SMOTE is used. Here, it is important to note that selecting between different models for encoding the data and imputing the missing data and oversampling data is not in the scope of this research.

4.4 Model selection:
The scope of this study has been made limited to the combined use of the SVM model and LDA model, the purpose of the study is not to select from different classification models but to increase the accuracy of the same classification model by solving the problem of imbalanced data and overlapping.

4.5 Evaluation criteria
The evaluation of this model is to be done by Area under the curve(AUC) value.

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \quad \ldots(1)$$

In, equation 1, TPrate is the rate of true positive in the confusion matrix and FPrate is the rate of false-positive in the final results. The AUC is a single measure of evaluating the classifying models (López, Fernández, García, Palade, & Herrera, 2013).

4.6 Training & prediction
The data training and testing split are done by stratified K-fold where 10 fold is used for all the models.

4.7 Hyperparameter tuning
The hyperparameter tuning of SVM has been done with the help of GridSearch in this research.

In conclusion, the selection of the methodology is being done on the hypothesis to reduce the number of false positives with SVM, LDA, and SMOTE function.

5. Results and discussion:

	Overall accuracy	AUC	Precision	Recall	F-score
SVM, LDA, and SMOTE	0.7	0.6904	0.8333	0.7142	0.7692
SVM & LDA	0.74	0.5761	0.7340	0.9857	0.8414
SVM	0.72	0.5523	0.7234	0.9714	0.8292

Table 1:The results of a different combination of SVM, LDA and SMOTE with overall accuracy, AUC, Precision, Recall, and F-score.

Here, the data in Table-1, shows the importance of methods of SVM, LDA and SMOTE, simultaneously. Here, it can be noticed that even though the overall accuracy of the second and third cases is higher than the overall accuracy of the first case, the AUC value for both is comparatively lower.

The results of table 1, supports our hypothesis, that after the problem of imbalance has been solved by the SMOTE function and the problem of overlapping has been reduced by increasing the class separability with LDA, the number of false-positive has been reduced. Also, the last case from table 1, about using the original data in SVM shows lesser accuracy due to the class imbalance, which is similar to (Liu, Yu, Huang, & An, 2011). Moreover, it can be further deduced that, just like SVM, the LDA's performance also decreases with the imbalanced class distribution.

In conclusion, the class separability of LDA decreases in imbalanced class distribution. And, after using both LDA and SMOTE at the same time the accuracy of the model in terms of AUC value increases significantly.

6. Conclusion:

The results of this study indicate improvement in the accuracy after applying SMOTE in the data and using LDA with SVM. It can be concluded from the results that by using SMOTE the machine learning model is being fed with more relevant minority class data. And due to getting more minority class data, the machine learning model can recognize minority class data more efficiently. Moreover, the use of LDA in data reduction can help separate the two-class clusters and also helps the SVM in selecting the most efficient hyperplane. Further study in this field is required as this particular problem of imbalanced class distribution is more common in real-life data.

7. References:

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi: 10.1613/jair.953

Ferreira, L. (2018, January 9). German Credit Risk - With Target. Retrieved from https://www.kaggle.com/kabure/german-credit-data-with-risk

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, *36*(2), 3302–3308. doi: 10.1016/j.eswa.2008.01.005

Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing & Management*, *47*(4), 617–631. doi: 10.1016/j.ipm.2010.11.007

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141. doi: 10.1016/j.ins.2013.07.007

Xiong, T., & Cherkassky, V. (n.d.). A combined SVM and LDA approach for classification. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* doi: 10.1109/ijcnn.2005.1556089

Ye, J. J. (2005). Two-dimensional linear discriminant analysis. *Advances in Neural Information Processing Systems*.