Sensitivities of Classification Algorithms to Over-sampling and Under-sampling Methods

Kairui Wu
Student Number: 1004503833
School of Information
University of Toronto
Toronto, ON, M5S 1A1
kairui.wu@mail.utoronto.ca

Abstract

When dealing with imbalanced datasets, classification algorithms may not work properly. Over-sampling and under-sampling methods are popular to be used to create a balanced dataset, for which the performance of the model will be enhanced. This project did the research on sensitivities of different classification algorithms to over-sampling and under-sampling methods when dealing with the imbalanced dataset. The results show that most of classification algorithms are more sensitive to under-sampling methods than to over-sampling methods. Besides, Logistic Regression is sensitive to both methods while Naïve Bayes is not influenced by sampling methods.

1 Introduction

An imbalanced dataset is a dataset of which the total number of a class of data is far less than the total number of another class of data. This is very common in practice such as fraud detection, load default detection, anomaly detection, etc. Classifiers are more sensitive to detect the majority class and less sensitive to the minority class. Thus, if we don't take care of the issue, the classification output will be biased, in many cases resulting in always predicting the majority class.

Over-sampling and under-sampling methods are useful in preprocessing imbalanced datasets. Over-sampling typically adds new samples with a minority class label while under-sampling deletes samples with a majority class label. However, classifiers have different sensitivities to different sampling methods. Doing the research on sensitivities of different classification algorithms to sampling methods is meaningful, since the conclusion can be provided as a reference to the application of classification algorithms on practical problems with imbalanced datasets.

This project used the dataset from the competition 'Give me some Credit' on Kaggle [1], of which the number of records with class-1 labels was only 6% of total records. The research method was applying SMOTE over-sampling method and Random Selection under-sampling method to change the imbalanced degree of the dataset to find changes of the value of AUC of models obtained by different classification algorithms. The change of the value of AUC illustrated the sensitivity of each algorithm to over-sampling and under-sampling methods.

2 Literature review

Several studies have been undertaken to compare performances of different classification algorithms when dealing with imbalanced datasets. For example, Brown and Mues compared performances of 10 classification algorithms when dealing with datasets of different imbalanced degrees in 2011 [2]. In their conclusion, random forest achieved the best performance in dealing with the imbalanced dataset. Logistic regression also had a good performance. Decision tree was significantly worse than others. However, they only used under-sampling methods to preprocess the data and did not focus on the sensitivity of the classification algorithm to sampling methods. In 2003, the research on the comparison of 17 classification algorithms on imbalanced datasets was finished by Baesens et al [3]. This research concluded that both complicated algorithms such as neural networks and SVM and simple linear algorithms such as linear discriminant analysis and logistic regression achieved very good performances with the sampling. However, some of researches got conflicting results. For example, for the research finished by Yobas, Crook, and Ross in 2000 [4], the conclusion presented that linear discriminant analysis had a better performance than neural networks in the prediction of loan default, whereas in the research finished by Desai, Crook, and Overstreet [5], the result was opposite.

Besides, various sampling techniques have been researched with classification problems. For example, Bastista applied 10 sampling techniques on 13 datasets in 2004 and found generally over-sampling methods provide more accurate results than under-sampling methods [6]. In 2013, Ramezankhani and Azizi did a research on the impact of over-sampling with SMOTE on the performance of neural networks, decision tree and naïve bayes [7]. The result illustrated that the best performances for decision tree and neural networks were achieved using totally balanced data but the performance of naïve bayes was not influenced by the sampling. However, they did not use the value of AUC as an indicator to evaluate the performance of the model and samples in their project were not heavily imbalanced.

3 Research design

3.1 Dataset

The dataset had 250,000 observations in total, which were provided by borrowers in the past years. The training set had 150,000 observations and the test set had 100,000 observations. This dataset contained some missing values. Around 21% of the observations had missing values. Also, most of the input features had outliers. Besides, in training dataset, only 6% of the records had a label of class-1, which indicated the characteristic of imbalance of this dataset. There were 10 independent variables in the dataset, all of which were real-number-based variables. The dependent variable was a binary variable, which indicated that detecting loan defaults is a binary classification problem.

3.2 Classification Algorithms

In this project, five classification algorithms were applied, which are shown in the Table 1.

Table 1: Algorithms

Algorithms	Logistic Regression	Random Forest	Decision Tree	KNN	Naïve Bayes
Settings	Scikit-learn Default	Scikit-learn Default	Scikit-learn Default	Scikit-learn Default	Scikit-learn Default

3.3 Sampling Methods

SMOTE algorithm is a popular over-sampling technique and achieves good performances in various researches [8]. While different techniques have been proposed in the past, typically using more advanced methods (e.g. under-sampling specific samples, for examples the ones “further away from the decision boundary” [9]) did not bring any improvement with respect to simply selecting samples at random. Therefore, in this project, SMOTE and Random Selection sampling methods were used to preprocess the dataset.

3.3.1 Over-sampling: Synthetic Minority Over-sampling Technique

SMOTE creates synthetic observations of the minority class (bad loans) by:

1) Finding the k-nearest-neighbors for minority class observations. (finding similar observations)
2) Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked, new observation.

3.3.2 Under-sampling: Random Selection Method

This method changes the class-1 label ratio through the following steps:

1) Selecting all majority class observations.
2) Randomly selecting observations that are kept based on the aimed ratio.
3) Concatenating selected observations and all minority class observations to be the new dataset.

3.4 Research Process

The research process of this project is shown in Figure 1. This process focused on sensitivities of classification algorithms on SMOTE and Random Selection sampling methods and compared the change of AUC values at different imbalanced degree realized by each sampling method.

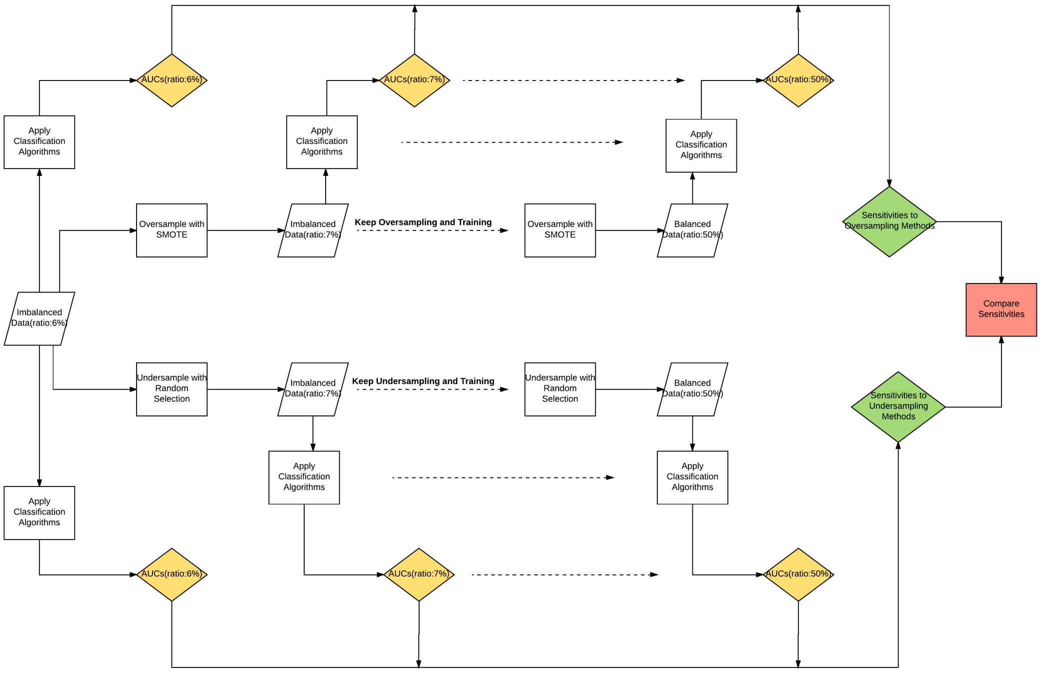
AUCs(ratio:6%)
AUCs(ratio:7%)
AUCs(ratio:50%)
Apply Classification Algorithms
Apply Classification Algorithms
Apply Classification Algorithms
Oversample with SMOTE
Imbalanced Data(ratio:7%)
Keep Oversampling and Training
Oversample with SMOTE
Balanced Data(ratio:50%)
Sensitivities to Oversampling Methods
Imbalanced Data(ratio:6%)
Compare Sensitivities
Undersample with Random Selection
Imbalanced Data(ratio:7%)
Keep Undersampling and Training
Undersample with Random Selection
Balanced Data(ratio:50%)
Sensitivities to Undersampling Methods
Apply Classification Algorithms
Apply Classification Algorithms
Apply Classification Algorithms
AUCs(ratio:6%)
AUCs(ratio:7%)
AUCs(ratio:50%)

Figure 1: Research Process Flow Chart

4 Conclusion

Most of algorithms used in this project presented apparent sensitivities to Random Selection under-sampling method, whereas there were only two algorithms which were sensitive to SMOTE. Logistic regression presented high sensitivities to both of sampling methods used in this project. Naïve bayes presented the best performance in dealing with imbalanced datasets but did not be influenced by the sampling.

4.1 ROC of classification algorithms on the original dataset

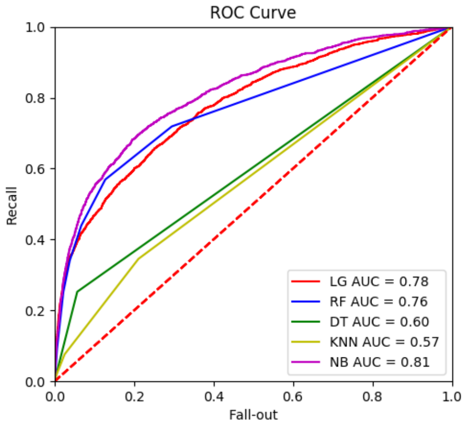

Figure 2: ROC on Original Dataset

At the imbalanced degree of 6% class-1 label ratio, naïve bayes algorithm achieved the best performance, whereas KNN and decision tree performed worst among these five classifiers. (Figure 2)

4.2 Sensitivities to SMOTE over-sampling method

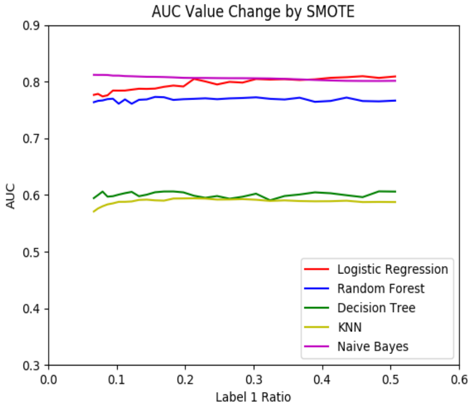

Figure 3: AUC Change by SMOTE

Table 2: AUC Change by SMOTE

Class-1 Label Ratio	**7%**	**8%**	**9%**	**9%**	**10%**	**11%**	**12%**	**13%**	**14%**	**16%**	**17%**	**18%**	**20%**	**21%**
Logistic Regression	0.78	0.77	0.78	0.78	0.78	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.80
Random forest	0.76	0.77	0.77	0.77	0.76	0.77	0.76	0.77	0.77	0.77	0.77	0.77	0.77	0.77
Decision tree	0.59	0.61	0.60	0.60	0.60	0.60	0.61	0.60	0.60	0.60	0.61	0.61	0.60	0.60
KNN	0.57	0.58	0.58	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59

Naïve Bayes	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81

Class-1 Label Ratio	**23%**	**25%**	**26%**	**28%**	**30%**	**32%**	**35%**	**37%**	**39%**	**41%**	**44%**	**46%**	**48%**	**51%**
Logistic Regression	0.80	0.79	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.81	0.81	0.81	0.81	0.81
Random forest	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.77	0.77	0.77	0.77	0.77
Decision tree	0.59	0.60	0.59	0.60	0.60	0.59	0.60	0.60	0.60	0.60	0.60	0.60	0.61	0.61
KNN	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
Naïve Bayes	0.81	0.81	0.81	0.81	0.81	0.81	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80

With SMOTE sampling method, most of algorithms were not sensitive to the change of the imbalanced degree. With the increasing of samples with class-1 labels, the performance of logistic regression kept being improved. The value of AUC increased from 0.78 to 0.81 with the ratio of samples with class-1 labels increasing from 7% to 50%. The performance of KNN was slightly improved on the same condition. The value of AUC obtained by KNN model increased from 0.57 to 0.59 with the ratio of samples with class-1 labels increasing from 7% to 50%. In addition, performances of naïve bayes, decision tree and random forest were not changed with the change of the imbalanced degree realized by SMOTE sampling method. (Figure 3 & Table 2)

4.3 Sensitivities to Random Selection under-sampling method

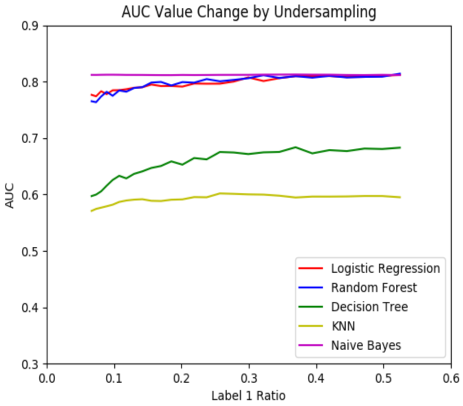

Figure 4: AUC Change by Random Selection

Table 3: AUC Change by Random Selection

Class-1 Label Ratio	**7%**	**8%**	**9%**	**10%**	**11%**	**12%**	**13%**	**14%**	**16%**	**17%**	**19%**	**20%**	**22%**
Logistic Regression	0.78	0.78	0.78	0.78	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.80
Random forest	0.77	0.77	0.78	0.77	0.78	0.78	0.79	0.79	0.80	0.80	0.79	0.80	0.80
Decision tree	0.60	0.61	0.62	0.63	0.63	0.63	0.64	0.64	0.65	0.65	0.66	0.65	0.66
KNN	0.57	0.58	0.58	0.58	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.60
Naïve Bayes	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81

Class-1 Label Ratio	24%	26%	28%	30%	32%	35%	37%	39%	42%	45%	47%	50%	52%
Logistic Regression	0.80	0.80	0.80	0.81	0.80	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Random forest	0.80	0.80	0.80	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Decision tree	0.66	0.68	0.67	0.67	0.67	0.68	0.68	0.67	0.68	0.68	0.68	0.68	0.68
KNN	0.59	0.60	0.60	0.60	0.60	0.60	0.59	0.60	0.60	0.60	0.60	0.60	0.60
Naïve Bayes	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81

With Random Selection under-sampling method, most of algorithms presented apparent sensitivities to the ratio of samples with class-1 labels. The most sensitive one was decision tree, the value of AUC was changed from 0.6 to 0.68 with the ratio of samples with class-1 labels changing from 7% to 50%. Radom forest and logistic regression also achieved apparent improvements of their performances on this condition. However, trends of their improvements became very small when the ratio of samples with class-1 label was larger than 30% and these three algorithms did not show any apparent improvement with the ratio changing from 30% to 50%. When the ratio was larger than 10%, KNN presented no change of its performance. Naïve bayes kept being the best and did not show any improvement with the sampling. (Figure 4 & Table 3)

5 Comparison

Random forest and logistic regression both achieved good performances when dealing with the imbalanced dataset with the under-sampling method. This point is well aligned with the conclusion of the research finished by Brown and Mues in 2011 [2]. Besides their conclusions, this project also illustrated the excellent performance of naïve bayes when dealing with imbalanced datasets and presented that KNN did not have such a high sensitivity to the under-sampling as claimed in their conclusion.

Logistic regression, decision tree and random forest presented much higher sensitivities to the change of the imbalanced degree realized by Random Selection under-sampling method than by SMOTE, for which they had worse performances with SMOTE. This conclusion is opposite to which claimed by Bastista [6].

Compared with the conclusion of the research finished by Ramezankhani and Azizi that decision tree had a good sensitivity to the change of the imbalanced degree realized by SMOTE [7], the conclusion of this project presented a low sensitivity of decision tree to SMOTE. In their research, they did not use AUC as an indicator to the model performance or either do research on heavily imbalanced datasets. Although the specificity and accuracy of decision tree were sensitive to SMOTE as claimed in their conclusions, this project presented that the performance of decision tree which could be indicated by the value of AUC was not improved by the sampling with SMOTE.

6 Limitation and further work

The parameters of all classification algorithms used in this project were set by default value in skicit-learn package in Python. However, the performance of the algorithm such as KNN, random forest, decision tree is highly influenced by the value of hyper-parameters. The default value of hyper-parameters may cause bad model performances, for which algorithms' actual sensitivities to sampling were not presented.

There are other sampling methods besides SMOTE and Random Selection. The algorithms' sensitivities to the imbalanced degree realized by various sampling methods will be different due to mechanisms of algorithms and sampling methods. Using more sampling methods to test algorithms' sensitivities will be more convincing.

Other than AUC, there are many other indicators that illustrate characteristics of models such as accuracy, F1 score and specificity. Using various indicators to illustrate algorithms' sensitivities will be more comprehensive.

Further work on this project can use the cross-validation method to find optimal hype-parameters to deal with this dataset, and then test sensitivities of classifiers. Besides, using various sampling methods and indicators will give us a deeper understanding on classification problems with different imbalanced degrees achieved by different sampling methods.

References

[1] http://www.kaggle.com/c/GiveMeSomeCredit

[2] Brown, I., Mues, C. (2011). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39 (2012), 3446–3453.

[3] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635.

[4] Yobas, M. B., Crook, J. N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. IMA Journal of Management Mathematics, 11(2), 111–125.

[5] Desai, V. S., Crook, J. N., & Overstreet, G. A. Jr., (1996). A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research, 95(1), 24–37.

[6] Batista, G. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29.

[7] Ramezankhani, A., Azizi, F. (2016). The Impact of Oversampling with SMOTE on the Performance of 3 Classifiers in Prediction of Type 2 Diabetes. Journal of MEDICAL DECISION MAKING, 2016(36), 137–144

[8] https://www.jair.org/media/953/live-953-2037-jair.pdf

[9] Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 200 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning Las Vegas, Nevada.