Exploring Different Approaches to Improve Binary Classification Performance for Imbalanced Data Set

Yuan Sun
Department of Computer Engineering
University of British Columbia
Vancouver, BC
anna9501@ece.ubc.ca

Yixuan Ji
Department of Computer Engineering
University of British Columbia
Vancouver, BC
jiyixuan@ece.ubc.ca

Jay Fu
Department of Computer Engineering
University of British Columbia
Vancouver, BC
jay.fu@alumni.ubc.ca

Abstract

This is a good abstract.

1 Introduction

This is instruction section.

1.1 sub

This is a good sub section.

1.2 sub

This is a good sub section.

2 Related Work

This is related work section.

3 Descriptions and Justifications

To boost up the accuracy of a model and minimize the effect of imbalanced data on the performance, there are several possible solutions, including undersampling, oversampling and class weighting. However, seldom have studied the difference in these approaches. The motivation of this experiment is to study how different machine learning techniques could have impact on the performance of models trained on imbalanced data set. The following sections discuss in detail the settings of the experiment and some of the approaches to tackle imbalanced problems.

3.1 Data Set Settings

The data set is drawn from the UCI Machine Learning Repository and is publicly known as the *Adult* data set (Kohavi and Becker, 1996). It contains general information of 48843 individuals

Table 1: *Adult* Data Set Attributes (Kohavi and Becker, 1996)

Name	Description
Age	Continuous
Workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Final Weight	The number of people the census believes the entry represents; Continuous
Education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education-Num	Continuous
Marital-Status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Female, Male
Capital-Gain	Continuous
Capital-Loss	Continuous
Hours-per-Week	Continuous
Native-Country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad& Tobago, Peru, Hong, Holand-Netherlands
Salary	>50K, <=50K

and whether or not they make more than 50K every year. The goal is to build a model that can accurately predict the annual income ($> 50K$ or $<= 50K$) of a given person based on this data set. As shown in Table 1, the data set contains a mixture of categorical and numerical features for each entry. Therefore, data pre-processing techniques should be applied to enable further study on the data. Specifically, feature selecion algorithms could help decide which features are relevant to the prediction. In addition, the label $salary$ is a binary attribute of the individual's income being either $> 50K$ or $<= 50K$, which makes it an ideal binary classification problems.

The major challenge for this data set is the imbalance of its binary label. There are only 11687 positive ($> 50K$) labels out of 48843 entries in total, which makes up 23.9% of the whole data set. Models trained on imbalanced data set tend to make prediction of the majority class. For example, consider a data set consisting 10000 entries of class A and 100 entries of class B. The model could get 90$\%$ of training accuracy by simply predicting everything as class A. The following sections discuss several methods to handle imbalanced data set, including previous efforts that has been made to study this data set, as well as other machine learning techniques that could also be applied.

3.2 Mearsure of Performance

In order to compare different algorithms, let us first define the measure of performance. In the sense of binary classification, predictions can be categorized into four different types: true positive (TP), false positive (FP), true negative (TN) and false negative (FP). In different applictions, performance can be measured using different formulae. The following three are the most used formulae (Zhou and Lai, 2009) in binary classification:

$$Sensitivity(SEN) = \frac{TP}{TP+FP} \tag{1}$$

$$Specificity(SPE) = \frac{TN}{TN+FN} \tag{2}$$

$$PredictiveAccuracy(PA) = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

$Sensitivity$ focuses on the model's accuracy on its positive predictions whereas $specificity$ focuses on negative predictions. $Predictiveaccuracy$ on the other hand, is used to measure the general accuracy of the model over both positive and negative predictions. In this experiments, we compare the performance of different algoritms using all three of the measures.

3.3 Undersampling

One of the most popular methods in handling imbalanced data sets is data undersampling, in which entries are randomly sampled from the majority class. Only a portion of the majority class are collected such that the size of the sampled majority class is the same as the minority class. Inouye (2018) has implemented such sampling method to the $Adult$ data set. He randomly generated a subset of sample with income $<= 50K$ and simply discarded the rest. This approach usually works relatively well, however, massive amount of data is discarded and the resulted model would not be able to reflect the comlete data set. Next section introduces an alternative method that could potentially utilize every entry of the data set to build the model.

3.4 Oversampling

As an alternative to the undersampling method, oversampling "creates" new entries of the minority class to match the size of majority class. The simplest way to oversample is to duplicate the minority class multiple times. However, there are more sophisticated oversampling algorithms, such as SMOTE and ADASYN, that can create synthetic data points as oppose to duplicating original entries. The particular algorithm used in this experiment is ADASYN, which randomly generates a data point based on k nearest neighbors. In addition, ADASYN attributes more weights to data points that are harder to learn (He et al., 2008). Both undersampling and oversampling are types of data pre-processing methods to solve imbalance problems. Next section introduces another technique that could be applied in the training phase of model fitting.

3.5 Class Weighting

Class weighting is a method that assigns more weights on important classes so that the model focuses more on one class than the other during training. This method can also be applied to imbalanced data set so the model would focus more on the minority class. Lo et al. (2008) applied class-balanced support vector machine (CB-SVM) to solve imbalanced problems. They assigned different weights to each class to prevent the model from favoring majority class. Similarly, class weighting is also conducted in this experiment using the svm module in $scikit-learn$ package. The specific function used, SVC, allows users to specify weights for each class through the parameter $class\_weight$.

3.6 Ensemble

Another approach that could potentially enhance the performance is model ensembling. In general it could help reduce the training error and/or approximation error of a variety of problems. However, seldom have applied ensemble method to the $adult$ data set. In this experiment, we apply ensemble method to reduce false predictions of the model. Specifically, we gather the outputs of a serie of heterogeneous machine learning models and train another logistic regression model based on these outputs. This is known as the stacking ensemble method. Section 4 discusses in detail how different techniques, including ensemble methods, would impact on the performance.

4 Experiments

We have done two experiments to study the effects of data balancing techniques and ensemble methods on performance. All hyper-parameters used in this section are determined using 10-fold

Table 2: Data Balancing Techniques

Algorithm	Undersampling			Oversampling		
	SEN	SPE	PA	SEN	SPE	PA
Random Forest	62.5	**92.6**	83.6	69.5	88.8	84.7
KNN	37.3	**83.5**	67.1	42.9	82.7	72.6
Decision Tree	56.9	**94.2**	81.3	62.5	90.2	83.0
Logistic regression	61.4	**92.5**	83.0	69.5	88.5	84.5
Neural Network	**73.6**	82.0	81.1	62.6	**91.3**	**83.4**
SVM	60.8	81.1	78.7	34.2	**81.6**	65.5

Algorithm	Class Weighting			None		
	SEN	SPE	PA	SEN	SPE	PA
Random Forest	**73.9**	88.1	**85.3**	71.1	88.5	84.9
KNN	N/A	N/A	N/A	**63.0**	81.1	**79.1**
Decision Tree	45.3	93.8	71.6	**77.3**	88.0	**86.1**
Logistic regression	**72.9**	88.2	**85.3**	72.6	88.0	85.0
Neural Network	N/A	N/A	N/A	N/A	76.3	76.3
SVM	58.9	81.1	78.4	**77.6**	80.6	**80.4**

cross validation. After shuffling the raw data set, we devide it into training set and testing set by a ratio of 7 : 3. All algorithms use the same training set and test set for comparison purposes. In addition, we perform one-hot encoding to all categorical features. In particular, for the missing categorical data, one-hot representation would simply be all zeros. And missing numerical data is filled with average of the corresponding feature column.

4.1 Data Balancing Techniques

The first experiment is to compare the performance of models after applying three different data balancing techniques as previously introduced: undersampling, oversampling and class weighting. In order to study only the effects of these techniques, we have controlled as many variables as possible. All three techniques use the same *adult* data set and same machine learning algorithms with the same hyper-parameters. In order to better observe how each technique influences the result, we conduct another control group in which no data balancing techniques are used. See Table 2 for the performance of each algorithms using different balancing techniques. As introduced in Section 3.2, binary classification models can be measured using SEN, SPE and PA. The corresponding scores are presented in the table 2.

As shown in table 2, undersampling method gets the highest specificity in most models. On the other hand, oversampling has better results in neural network model. As neural network models usually has the highest complexity, theoritically it does need more training data. This might explain why neural network would work better in oversampling, because it greatly increases the amount of data. Class weighting could improve the performance in random forest and logistic regression, but it failed to converge for KNN and neural network models. As the nature of class weighting method, it changes the weight of classes and thus the gradient in each iteration. This might be one of the reasons for the non-convergence. Also finding the right weights is challenging as the training generally takes very long. It is likely that the models could have converged or even achieved better performance if better set of weights were used. Lastly, the control group without using any balancing techniques did surprisingly well and even outperformed others in some of the measurements. This provides an insight that applying balancing techniques might not always be beneficial. When dealing with imbalanced data set, it is possible to simply train on the raw data and get better results than applying some balancing method beforehand.

Table 3: Ensemble vs. Single Models

Model	SEN	SPE	PA
Stacking Ensemble	**69.7**	**86.7**	**83.3**
Random Forest	62.5	92.6	83.6
KNN	37.3	83.5	67.1
Decision Tree	56.9	94.2	81.3
Logistic Regression	61.4	92.5	83.0
Neural Network	73.6	82.0	81.1
SVM	60.8	81.1	78.7

4.2 Ensemble Method

The second experiment explores how ensemble method can affect the model performance. From the result of Section 4.1, we determine that oversampling performs better in most of the algorithms in this binary data set. Therefore Over Sampling is used in both groups of this experiment in order to control experimental factors. Similar to the previous part, models are compared in three different matrices SEN, SPE and PA.

As shown in table 3, ensemble method could achieve relatively high performance among all the single models. Through the stacking process, the model is able to optimize its performance to be close to the best single model in each measurement. However, it does not seem to have a huge advantage over other models. As a future project, we could try more ensemble methods to further study the effect of ensembling in imbalanced data.

5 Discussion

In this experiment we have applied different data balancing techniques and ensemble method to the imbalanced $adult$ data set. It is found that undersampling usually performs well in most models while oversampling tends to do better with neural network. Class weighting technique could potentially inprove performance but it is not as stable as other methods. Interestingly, balancing techniques do not neccessarily result in higher performance. For this data set, model that did not apply any balancing achieved higher score in some measurements.

In addition, stacking ensemble is able to get relatively high performance among all the models. This is because when training the stack model it would tend to have more "trust" in the model that makes the right prediction. In other words, the resulting stack model would perform nearly as well as the best single model.

One of the major contribution of this experiment is that our work has provided some insights to the performance of different balancing techniques. It could serve as a reference for any further project or research related to imbalanced data set. The results might help researchers make decision in what machine learning techniques should be applied to their particular imbalanced problems.

In the ensemble experiment, we only used oversampling in both groups to control experimental factors. However this might neglect the possibility that a different balancing technique might make a difference in the ensemble results. If time permits, one possible improvement is to conduct ensembling to all balancing techniques and study their difference. In addtion, we could try more ensembling method besides stacking, especially those that could reduce training error.

Another potential future improvement is to expand the experiment to more data sets that have different imbalance ratio. This could help explore how balancing techniques and ensemble methods would perform differently with different imbalance ratio.

References

H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *Annalen der Physik*, pages 891–921, 2008.

S. Inouye. Census income classification in r. *Inertia 7*, 2018. URL `https://www.inertia7.com/projects/146`.

R. Kohavi and B. Becker. Adult data set, 1996. URL `https://archive.ics.uci.edu/ml/datasets/Adult`.

H.-Y. Lo, C.-M. Chang, T.-H. Chiang, C.-Y. Hsiao, A. Huang, T.-T. Kuo, W.-C. Lai, M.-H. Yang, J.-J. Yeh, C.-C. Yen, and S.-D. Lin. Learning to improve area-under-froc for imbalanced medical data classification using an ensemble method. *SIGKDD Explorations*, 10(2):43–46, 2008.

L. Zhou and K. K. Lai. Benchmarking binary classification models on data sets with different degrees of imbalance. *Front. Comput. Sci. China*, 3(2):205–216, 2009. doi: http://dx.doi.org/10.1007/s11704-009-0027-1.