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

Yuan Sun
Department of Computer Engineering
University of British Columbia
Vancouver, BC
round.sun@alumni.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. It contains general information of 48843 individuals and whether or

Table 1: Data Set Attributes

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

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. Then 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 tackle this problem in particular, as well as additional machne learning techniques that could also be applied.

3.2 Mearsure of Performance

In order to compare different algorithms, it is necessary to 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 in binary classification:

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

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

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

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

3.3 Undersampling

One of the most popular methods to handle 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 to randomly generate 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 fully 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. There are, however, 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 with more weights are attributed to data points that are harder to learn. Both undersampling and oversampling are types of data pre-processing methods 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 samples with 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 SVM (CB-SVM) to solve imbalanced problems, in which different weights are assigned to each class as an effort of preventing 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, provides interface that allows users to specify weights for each class through the parameter $class\_weight$, which has greatly facilitated this experiment.

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 problems in a variety of domains, rather than just imbalanced data set. However, seldom have applied ensemble method to the $adult$ data set. In this experiment, ensemble methods are used as an attempt to reduce false predictions of the model. Specifically, stacking ensemble method is used in which outputs of a series of heterogeneous machine learning models are gathered as inputs to train a logistic regression model. Predictions are made from the output of this stacking model. Section 4 discusses in detail how different techniques, including ensemble methods, would impact on the performance.

4 Experiments

This is experiments section.

4.1 sub

This is a good sub section.

4.2 sub

This is a good sub section.

4.3 sub

This is a good sub section.

4.4 sub

This is a good sub section.

5 Discussion

This is discussion section.

5.1 sub

This is a good sub section.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can use more than eight pages as long as the additional pages contain *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.