Journal of Cases on Information Technology
Volume 19 • Issue 1 • January-March 2017

Improving Classification Accuracy on Imbalanced Data by Ensembling Technique

Divya Agrawal, Shri Shankaracharya College of Engineering and Technology, Bhilai, India
Padma Bonde, Shri Shankaracharya College of Engineering and Technology, Bhilai, India

ABSTRACT

Prediction using classification techniques is one of the fundamental feature widely applied in various fields. Classification accuracy is still a great challenge due to data imbalance problem. The increased volume of data is also posing a challenge for data handling and prediction, particularly when technology is used as the interface between customers and the company. As the data imbalance increases it directly affects the classification accuracy of the entire system. AUC (area under the curve) and lift proved to be good evaluation metrics. Classification techniques help to improve classification accuracy, but in case of imbalanced dataset classification accuracy does not predict well and other techniques, such as oversampling needs to be resorted. Paper presented Voting based ensembling technique to improve classification accuracy in case of imbalanced data. The voting based ensemble is based on taking the votes on the best class obtained by the three classification techniques, namely, Logistics Regression, Classification Trees and Discriminant Analysis. The observed result revealed improvement in classification accuracy by using voting ensembling technique.

KEYWORDS

Classification, Ensembling, Logistic Regression, Telemarketing

1. INTRODUCTION

Marketing selling campaign uses a typical strategy to enhance the business where direct marketing is one of the easiest approach which eases the direct marketing. As the application area of the technology increases the data also increases. Classification of data becomes difficult because of unbounded size and nature of the data. Class imbalance problem becomes greatest issue in data mining. This technology basically focuses on increasing customer lifetime value by using customer metrics. Mainly the task is to select the best no. of clients. Data mining technique plays a key role in personal and intelligence DSSs, allowing the semi-automatic extraction of explanatory and predictive knowledge from raw data. In particular classification is the most common data mining task and the goal is to build a data driven model that learn an unknown underlying function that maps several input variables (Moro, Cortez and Rita, 2014). There are several classification models such as the logistic Regression (LR), Classification tree and Discriminant Analysis(DT). Logistic Regression(LR) and Classification Trees(CTs) are basically easily understandable by humans by easily fitting into the models and they also provide better prediction in classification task. After comparing with all these three models it

DOI: 10.4018/jcit.2017010104

shows different classification accuracy which still is a challenge to improve. So, in order to maximize the performance of classification accuracy introduce a voting based ensembling technique.

Classification accuracy is still a great challenge due to data imbalance problem. The increased volume of data is also posing a challenge for data handling and prediction, particularly when technology is used as the interface between customers and the company. As the data imbalance increases it directly affects the classification accuracy of the entire system. Area under the curve (AUC) and lift prove to be good evaluation metrics. AUC does not depend on a threshold, and is therefore a better overall evaluation metric compared to accuracy. Lift is related to accuracy and is widely well used in marketing (Burez and Van Den Poel, 2009). So, by using better metrics imbalance problems can be handle properly. Another way to improve classification accuracy is oversampling whereby, the training data set is randomly selected from both the classes and joined to form the training set. The rest is used as test / validation set. Thus, in effect the higher class is oversampled and the imbalance is removed. However, oversampling is criticized for changing the proportion of classes in the dataset.

Several classification techniques are in vogue such as Logistics Regression, Classification Trees and Discriminant Analysis. Some of the fields where classification techniques find application are Engineering, Finance, and Marketing. For example, a bank would want to predict the possibility of default on part of the customer before disbursing loan to him. Similarly, a company would want to predict the possibility of success before marketing a product in a certain area. However, one of the issues in the datasets used for prediction is that they are imbalanced. For example, in a dataset of 1000 loan disbursed, one may find 100 cases of defaults. Although, in 90% of cases in such situations there was no default, the rest 10% cases constitute tremendous loss for banks.

This mechanism proposed a voting based ensemble to improve classification accuracy. Ensemble Learning is a two-step decision making process, in which the first step is related to the decision of the individual classifier and the second step refers to the decision of the combined model. The idea behind ensemble methodology is to build a predictive model by voting on classes predicted by various classification techniques. It is well-known that ensemble methods can be used for improving prediction performance.

2. LITERATURE REVIEW

Data prediction in bank telemarketing has been an active research topic in the statistics, database, and security communities for the last three decades. As the companies are using direct marketing i.e., communicating with customers through various channels. Data mining basically allows rethinking marketing by focusing on maximizing customer lifetime value through the evaluation of available information and customer metrics. There are several classification models, such as the Logistic Regression (LR), Classification trees (CTs) and the more recent neural networks(NNs) and support vector machines (SVMs). LR and DT have the advantage of fitting models that tend to be easily understood by humans, while also providing good predictions in classification tasks (Moro, Cortez and Rita, 2014). The ROC is widely used as a measure of performance of classification rules. However, it has observed that the measure is fundamentally incoherent, in the sense that it treats the relative severities of misclassifications differently when different classifiers are used (Hand and Anagostopoulos, 2013).

Area under the curve AUC and lift proved to be good evaluation metrics. AUC does not depend on a threshold, and is therefore a better overall evaluation metric compared to accuracy. Lift is very much related to accuracy, but has the advantage of being well used in marketing practice (Burez and Van Den Poel, 2009). It is revealed that using sophisticated sampling techniques did not give any clear advantage. Weighted random forests, as a cost-sensitive learner, performs significantly better compared to random forests, and is therefore advised. It should, however, always be compared to logistic regression. Boosting is a very robust classifier, but never outperforms any other technique