2.3.3 C2)

RESEARCH ARTICLE

Adv. Sci. Lett. 4, 400-407, 2016

Advanced Science Letters
Vol. 4, 400-407, 2016

AMERICAN SCIENTIFIC PUBLISHERS

Imbalance classification model for churn prediction

Dech Thammasiri[1,3], Supoj Hengpraprohm[2,3], Kairung Hengpraprohm[2,3]and Suvimol Mukviboonchai[3]

[1]Faculty of Management Science, Nakhon Pathom Rajabhat University, Nakhon Pathom 73000, Thailand
[2]Faculty of Science and Technology, Nakhon Pathom Rajabhat University, Nakhon Pathom 73000, Thailand
[3]Machine Intelligence Research Unit, Nakhon Pathom Rajabhat University, Nakhon Pathom 73000, Thailand

Churn prediction deals with challenging problem of detecting customers who probably cancel a subscription to a service. Data mining techniques such as decision tree, logistic regression, neural network are very successful in prediction customer churn. However, the prediction accuracy of these classification techniques reduces when handling with class-imbalanced data. Class-imbalanced data are common in the field of Churn prediction, mainly one or some of the classes have much more instances samples in comparison to the others. Classification techniques for imbalanced datasets usually correctly predict the results for the majority class, but do not perform well to predict results for the minority class. In this paper, we propose SMOTEBagging, which combines SMOTE sampling technique with Bagging algorithm to enhance the classification model to predict results for the minority class. The classification performance is obtained via 5-fold cross validation. The experimental results show the effectiveness of SMOTEBagging technique.

Keywords: Churn prediction, Class imbalance, SMOTEBagging, Ensemble

1. INTRODUCTION

Telecommunication companies face with severe competition in this decade. They need to adjust themselves to survive in business. One of the strategies adopted by these companies is churn prediction to anticipate potential customers who have prepared to leave the companies. Using this technique, they can deliver special offers to maintain this group of customers. As a result, the companies can save costs of acquiring new customers to replace missing customers. There are many researchers presenting prediction models using machine learning and data mining to study churn prediction. Sharma, A. and P.K. Panigrahi[1] proposed Neural Network to perform predictive model. Moeyersoms, J. and D. Martens [2] conducted a study which used C4.5, Logit and SVM to study churn prediction in energy sector. Nie, G., et al.[3] developed forecasting credit card churn model using the logistic regression and decision tree. This previous review of studies shows that data mining methods give a high prediction accuracy rate.

However, there was a problem in the prediction that occurs due to the disproportion between churn and non-churn samples. Non-churn sample outnumbers churn sample and the two groups hold different characteristics. We call this problem "imbalance data problem". Imbalance data imposes problems because in creating a model, the algorithm learning process will focus more on the larger group while the smaller group possesses more important information. Many researcher proposes techniques to balance class distribution such as Burez, J., and Van den Poel, D. [4] proposed Under sampling and Over sampling technique to achieve the good accuracy in customer churn prediction. According to Thammasiri, D., et al.[5] used SMOTE to resampling for balance data. Their results showed that when using SMOTE data-balancing technique and SVM to classify data can achieved the good accuracy.

In this paper, the purpose of this research is to develop a classification model of persistence with maximum predictability. This paper is organized as follows. In Section 2, we explain the classification model such as Decision tree(DT), Neural Network(ANN), Support Vector Machine(SVM) and SMOTEBagging algorithms. In Section 3, we present the dataset, methodology and evaluation model by using confusion matrix used for our study. Section 4 we show experimental result. Finally, Section 5 concludes this research.

 1936-6612/2011/4/400/008 doi:10.1166/asl.2011.1261