HANDLING IMBALANCED DATA IN CHURN PREDICTION USING COMBINED SAMPLING AND WEIGHTED RANDOM FOREST

Veronikha Effendy[1], Adiwijaya[2], Zk. Abdurahman Baizal[3]

[1]Magister Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Pelanggan merupakan ase t utama dalam sebuah perusahaa n, tidak terkecuali industri telekomunikasi. Pelanggan churn merupakan masalah utama yang ditemukan dalam industri telekomunikasi , karena sangat berpengaruh terhadap pendapatan perusahaan. Pada saat proses churn mulai terjadi, biasanya persentase data yang menggambarkan pelanggan churn tidak banyak. Kondisi ini menyebabkan diperlukannya model prediksi untuk dapat menentukan pelanggan yang berpotensi melakukan churn. Pendekatan Data mining dapat menghasilkan model prediksi dengan mempelajari data historis transaksi pelanggan. Minimnya data pelanggan c hurn di antara sejumlah data pelanggan yang dimiliki oleh perusahaan akan menimbulkan masalah data yang tidak seimbang. Data yang tidak seimbang akan menyebabkan kesulitan dalam pembuatan model prediksi sehingga hasil prediksi pelanggan churn menjadi tida k akurat. Sampel data yang digunakan dalam penelitian ini memiliki persentase churn 0,7 %. Penelitian ini menerapkan teknik kombinasi sampling dan Weighted Random Forest (WRF) untuk menghasilkan model prediksi pelanggan churn. WRF telah dikembangkan dari m etode Random Forest (RF) untuk mengatasi masalah data yang tidak seimbang yang biasa terjadi pada prediksi churn. Metode ini diklaim dapat menghasilkan kinerja yang cukup baik pada data yang tidak seimbang. Namun, pada penelitian ini ditemukan kendala bahw a performansi yang dihasilkan masih rendah. Dengan menggunakan teknik sampling , permasalahan performansi yang masih rendah dapat diatasi. Adapun teknik sampling yang digunakan adalah simple under sampling dan Synthetic Minority Over - sampling Technique (SMO TE). Hasil penelitian ini menunjukkan bahwa kombinasi SMOTE dan simple under sampling terbukti dapat meningkatkan kinerja model prediksi yang dihasilkan oleh WRF.

Kata Kunci : Churn , Pred iksi , Metode Random Forest , Metode Weighted Random Forest, kombinasi sampling, SMOTE

Abstract

Informatics Engineering 2014 iii | P a g e ABSTRACT Customers are the key asset in an industry , the telecommunications industry is no exception . Customer churn is a major problem that is found in the telecommunications industry , because it affect s the company's revenue. At the time of the customer churn is taking place, the percentage of data that describes the customer churn is usually not much , unfortunately the churn data is the data which have to predict earlier. This condition causes the need for predictive models in order to determine the potential customers do churn. D ata mining approach can produce prediction models by studying the historical data of customer transactions. The lack of data on customer churn among a number of customer data held by the company will lead to the problem of imbalanced data . Data that is not balanced will cause difficulty in making a prediction model so that the results of c ustomer churn prediction become inaccurate. The sample data used in this study has a percentage of 0.7 % churn. This research applies a combination of sampling techniques and Weighted Random Forest (WRF) to produce the customer churn prediction model . WRF has been developed from the method Random Forest (RF) to overcome the problem of unbalanced data which is common in churn prediction. This method is claimed to produce a reasonably good performance on the imbalanced data. However, this study found that the performance of the result s is still low. By using sampling technique , the low performance problems can be overcome. The sampling tech nique used is simple under sampling and Synthetic Minor ity Over - sampling Technique (SMOTE) . The results of this study indicate that the combination of SMOTE and under - sampling simple proven to increase performance of predictive models generated by WRF .

Keywords : Keywords: Churn , Prediction, Random Forest , Weighted Random Forest, Combined - sampling, SMOTE

CHAPTER 1: THE PROBLEM

This research was dealing with the problem of handling imbalanced data in churn prediction. This section discusses the rationale, theoretical framework, conceptual framework/paradigm, statement of the problem, hypothesis, assumption, scope and delimitation, importance of the study.

1.1 Rationale

Nowadays, telecommunication industries have a problem, it is concerning with customer churn, because this can affect the company's revenue. [1]. To survive and win the market competition, some companies attempt to predict customer churn with data mining approach [2].

Data mining approaches can help a company understand customer behavior from its own data, so that the company can implement the right CRM (Customer Relationship Management) strategies in order to save its revenue [2]. Unfortunately, churn is rare objects, the data of churned customers are only a few; however, it is of great interest and valuable for a company [3]. In other word, the data set for this case are extremely imbalanced.

Researchers have attempted to find methods to handle the imbalanced data in churn prediction. Some of them focuses on the data pre-processing, i.e. oversampling [4], and the other try to find the match classifier for this kind of problem, such as : logistic regression, linear classification, naïve Bayes, decision tree, multilayer perceptron neural networks, support vector machine, data mining evolutionary algorithm [1], and random forest [5], [6]. Some of those researches have resulted a good performance in churn prediction for their own data set, however every data set have their own characteristic and specific case [2].

There are two common approaches in handling imbalanced data. First is sampling approach and second is cost-sensitive approach [6]. There are continuous researches in improving prediction performance for handling imbalanced data using random forest, such as balanced random forest, weighted random forest [6], improve balanced random forest [1], weight random forest with under sampling [5], etc.

Weight random forest classifier claim to handle imbalanced data with cost-sensitive approach, that is to assign weight, so that it can reduce misclassified data [6] .

Sampling basic techniques are under sampling and oversampling. Each technique has its own benefits and drawbacks. Under sampling makes the model run faster, but this technique causes big loss of potential data from the majority class in imbalanced data, so that it reduces prediction performance. Oversampling create additional data (but not additional information), this causes slower running process. Although oversampling does not reduce the data record, but the additional data from the minority class causes over fitting (there are any possibilities that sampling makes any data in majority class moves to minority class) [4].

The data set used in this research is customer behavior profile data. It has been studied before in Indonesia. One of the results shows that SMOTE (Synthetic Minority Over-Sampling) algorithm has a good performance [4], however, the prediction measurement is not accurate. This research try to predict whether customers potentially churn or not-churn based on customer behavior profile with a high accuracy.

The basic idea in SMOTE is to create data synthetic from minority class [7]. However, SMOTE has a drawback, it makes over fit data. The combination of simple under sampling and SMOTE algorithm; may reduce the substantial loss of potential data from majority class and also reduce the probability of over fitting problem.

For those reasons, this research attempts to combine the two approaches (sampling and cost sensitive-learning) by applying dataset processed in combine sampling method (SMOTE for minority data and simple under sampling for the majority data) to the Weight Random Forest classifier [7], [6], [8].

1.2 Theoretical Framework

This research attempts to handle the imbalanced data in churn prediction. Data input for the system was a dataset containing customer profile from a specific product in a telecommunication industry in Indonesia.

The output of this system is churn prediction result and some results of performance measurement of the predictive model. The dataset used as input data in the system will go through the stages of pre-processing prior to produce clean data and can be recognized by the system, which is hereinafter referred to as the original data. Then, by using the tools WEKA, several variations of the input dataset created from the original data using the simple technique of under sampling [8], SMOTE [7] and a combination of both.

The process will be done in the system are as follows: 1) The data input divided into several parts and set up into training data and test data for 10-fold cross validation purposes [8]; 2) The process of forming multiple decision trees into a random forest involves data bootstrap, random attributes selection, weighted Gini criterion calculation, and some calculations to determine the label of each leaf node [6]; 3) The formed model was tested using test data that has been developed by the system; 4) The final prediction was the aggregating results from all of trees in the forest [6]; 5) After the final prediction was obtained, the system will calculate the performance in the 10-fold cross validation, and outputs the average performance of the overall validation [8].

1.3 Conceptual Framework/Paradigm

There are three variables applied to conduct measurement on this research, namely:

Variable	Variable's Information
Data composition	Various data composition is compiled from the original data which is manipulated using sampling techniques (i.e. SMOTE oversampling and simple under sampling)
WRF parameters	The parameter to induce WRF, consists of number of attributes that will be used in each single random tree, weight on each class, and number of tree induced in the forest.

1.4 Statement of the Problem

The main problem discussed in this research was handling imbalanced data in churn prediction by applying Weighted Random Forest (WRF) with sampling technique to improve the prediction performance.

Customer is a very important asset in an industry, particularly the telecommunications industry. The occurrence of customer churn can lead to a decrease in revenue of a company. Customer churn prediction is a prediction that is very important in an industry to be able to detect the potential customers will churn, so that churn can be prevented early. Because the data of customer churn is usually a minority of data from all the data that is owned by the company, there are difficulties in studying their characteristics. Various researches have been done until present time to develop a method that can address the case of imbalanced data, one of them is the ensemble method: weighted random forest [6]. This method created many classification trees and utilizes a greater weighting on the minority data, thus it can handle the imbalanced data [6] . However, for the case of the data set in this research (in which there are only 0.7 % of data churn), the previous research still cannot produce a good performance. Since there is an extreme imbalanced data, this research tried to implement combined-sampling techniques (combination of simple undersampling and SMOTE) and apply the WRF to be able to improve the performance of the prediction of customer churn [7], [8], [4].

1.5 Hypothesis

There are two common approaches in handling imbalanced data: first is sampling approach and second is cost-sensitive approach. This research attempts to implement WRF with combined-sampling technique to improve churn prediction performance. WRF is selected as the classifier for solving the cost-sensitive approach, because this classifier can handle imbalanced data and has a better true positive rate [6], while for the sampling approach simple under sampling technique is used; and it is combined with oversampling technique using SMOTE algorithm [7], [8]. This combination is used to avoid the big loss of the potential information in majority class. The combination of combined-sampling and WRF classifier are used to achieve the purpose of this research, i.e. the improvement of performance in true churn prediction.

1.6 Assumption

Global problem in churn prediction includes the variation of dataset, the churn prediction accuracy, the main factors causing churn, and the relation with marketing management to determine appropriate strategies to deal with the problem of churn. Continuous researches are needed to address this problem. This research focused on the problem of churn prediction accuracy and using the following assumptions:

1. This research was conducted based on data churn in a specific product of a telecommunication company in Indonesia.
2. This research discusses issues overcome imbalanced data on the prediction of churn in order to produce good performance.
3. This research does not discuss more about the factors which most influence on the churn.

1.7 Scope and Delimitation

There are many methods and techniques which can be used to improve the performance of the churn prediction. In order to get more focused analysis on churn prediction performance resulted by the selected method, this research used scopes and delimitations as follows:

1. This research attempts to compare the performance of a churn prediction using WRF with sampling techniques and churn prediction performance using the WRF without sampling techniques.

2. This research does not discuss more about how SMOTE work. SMOTE is carried out using Weka.
3. The specification of the hardware used in this research is varied, so elapsed time is not taken into the analytical process.

1.8 Importance of the Study

This research deals with how the performance of WRF on the formation of a churn prediction modeling of the data set which is extremely imbalance (i.e. the data have the churn percentage below 1 %), and whether the sampling technique used in data pre - processing improve the prediction performance significantly or not.

If it is proven that the application of sampling techniques to improve churn prediction performance is quite significant, the result will be validated using several different datasets in further reasearch, so that the two approaches can be combined into one method that can handle the extremely imbalanced data.

The performance of the churn prediction model will also be measured using F-measure and top-decile. Top-decile is commonly used in the churn prediction and it is the important value in the marketing management level. If the generated model results high top-decile, then the model can be considered to be applied in the company that wants to improve its services to customers with appropriate and suppress the occurrence of churn in its products.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

This chapter presents conclusions of the research and recommendations for future work.

5.1 Conclusions

Combined-sampling (oversampling and undersampling) improved the F-measure value of the churn prediction model produced by WRF. Moreover, the implementation of undersampling reduced the number of data records in the dataset caused by the oversampling, so it minimized the computational cost.

The higher churn percentage on the dataset, the lower weight of churn class needed to obtain the best F-measure.

The higher the F-measure value, the higher the value weighted accuracy. However, the high value of F-measure and weighted accuracy do not guarantee that the value of top-decile will be high anyway. In terms of pure classification, the best classifier is a classifier which has the highest F-measure and highest accuracy. But in the case of churn prediction, in terms of management, related to the limited budget, the classifier which has a high top-decile with a reasonable performance might be the choice.

5.2 Recommendations

Further experiments using different dataset which has more or different attributes are needed to validate the result of this research. It may makes clear what factor influence the ntree in order to obtained the best F-measure.

Based on the results achieved in this research, the performance generated by the modeling is quite good, but it needs more research to be able to increase the value of top-decile. Churn prediction model wihch has a good performance and a good top-decile, can reduce the cost of doing treatment against potential customer churn. In order to help the management to determine the appropriate strategies, besides getting the right target (by increasing the value of top-decile), other work is necessary conducted to get what most attributes can influence customers to churn. Based on this information, management is expected to determine the appropriate action to be made to the appropriate target customers anyway. In the end, after the determined strategy is implemented, it is necessary to evaluate whether the big problems of churn can be resolved well or not.

REFERENCES

[1] Y. Xie, X. Li, E. Ngai dan W. Ying, “Customer churn prediction using improved balanced random forests,” *Elsevier,Expert System with Application 36,* pp. 5445-5449, 2009.

[2] D. M. Maharaj, “Evaluating Customer Relations in The Cell phone Industry,” *IJBSM (International Journal for Business, Srategy & Management) Vol 1 No1* , 2011.

[3] B. Huang, M. T. Kechadi dan B. Buckley, “Customer Churn Prediction in Telecomunications,” *Elsevier, Expert Systems with Applications,* pp. 1414-1425, 2012.

[4] Z. A. Baizal, M. A. Bijaksana dan A. S. Sastrawan, “Analisis Pengaruh Metode Over Sampling dalam Churn Prediction untuk perusahaan Telekomunikasi,” *SNATI ISSN: 1907-5022,* 2009.

[5] J. Burez dan D. d. Poel, “Handling Class Imbalance in Customer Churn Prediction,” *Elsevier.Expert Systems with Applications 36,* p. 4626–4636, 2009.

[6] C. Chen, A. Liaw and L. Breimenn, "Using Random Forest to Learn Imbalanced Data," Statistics Department of University of California, Berkeley, 2004.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[8] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Boston: Person Education,Inc., 2006.

[9] L. Breimann, Random Forest, Netherlands: Kluwer Academic Publishers, 2001.

[10] S. A. Neslin, S. Gupta, W. kamakura, J. Lu and C. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *Journal of Marketing Research,* pp. 204-211, 2006.

[11] R. Mattison, The Telco Churn Management Handbook, Oakwood Hills, Illinois: XiT Press, 2005.