CHAPTER 1

INTRODUCTION

The most important problem faced by large companies is the occurrence of customer churn, many companies have a very significant loss when loyal customers move to another company, so various ways are done, so that customer churn does not happen and keep the company's revenue does not suffer significant losses.

1.1 Rationale

Customers are the essential part of business in the telecommunication industries because customers are the main source of revenue for the company to run the business. This results in a customer churn or a customer's condition moving from one service provider to another [1] and this becomes an important issue that must be solved because it gives effect to lower the company's revenue. Therefore, customer churn prediction research is done by classification technique classify potential customers to churn, so for the final result will show whether a customer will churn in the future, so that the company can take steps to prevent it.

Some classification algorithms have been applied to churn prediction problems such as logistic regression, naive bayes, decision trees, artificial neural networks, and support vector machines, but individual classifier cannot achieve accurate results [2], besides large learning methods such as ANN and SVM have been confirmed to have a great noise. Therefore, the Random Forest ensemble technique introduced by Breiman (2001) [3] has been an effective and popular algorithm because of its relatively good classification performance and ease of use, but random forest does not take into account class imbalances which can degrade the performance of classification and increases the bias against the majority class [4].Thus it is important to treat imbalance data. In this study, using churn data that has the characteristics of imbalance class or customers who do churn very little with percentage usually 2% of the sample data [5].

Several studies have used some methods to overcome this when using Random Forest, such as (Chao Chen, 2013) [6] and (Yaya, Xiu, Ngai, Weiyun, 2008) [7] who apply the sampling technique (undersampling) for balance class and cost-sensitive learning for classification, Although the sampling technique can work well, cost-sensitive learning makes it more susceptible to noise, and the WRF process takes a lot of time. Time computing is not efficient. Besides the research done by (Veronikha dkk, 2014)[8] performs a data approach technique by combining two methods (undersampling and SMOTE) to handle the problem of imbalanced data on weighted Random Forest, but this research produces a low predictive model of performance and gives rise to a very large bias that affects the

accuracy of the prediction model, and the process is run separately between the data balance process and the predicted algorithm process, thus it cannot fully see whether there is the influence of balance data directly to the weighted random forest the algorithm

In this research, handle imbalanced data with data and algorithm approach called Balance Random Forest (BRF). Balanced Random Forest combine sampling technique and ensemble idea that implements under sampling to majority class. This is done because the minority class gives direct impact to performance measurement so that undersampling is better done than oversampling, but downsampling allows the loss of information because many parts of majority class are not used, so it needs balance downsampling, where balance downsampling is applied to BRF. Balance random forest is believed to be better in regards to its performance than Weighted Random Forest [7] In adsition, BRF is computationally more efficient with large imbalanced data.

However, Balance Random Forest also has deficiencies, in the process of balance random forest applying random undersampling and that have a drawback that some important discriminating instance may be discarded [9] and they make BRF have a little effect classification process learning decision tree thus influencing the classifier is generated [7].

For those reasons, this research will carry out a prediction of customer churn, with data imbalance customer churn telecommunication industry using Balanced Random Forest Algorithm. To improve performance on the classification results, this study will note the undersampling procedure to compose better training set from the available data by the undersampling strategy based on clustering [9] to the process Balanced Radom Forest algorithm; thus overall it is called Modified Balance Random Forest (MBRF).

1.2 Theoretical Framework

This research attempts to handle the imbalanced data in churn prediction. Data input for the system is a dataset containing customer profile from a specific product in 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 is real data set.

The dataset used as input data is divided into training data and test data. Training data sets and Testing Dataset contains a proportion of churners that is the representative of the actual population to approximate the predictive performance in a real-life situation [7]. This study constructs all variables in the same way for each dataset.

Churn prediction is solved by classification technique when the classification technique produces a classifier. This study, Random Forest is used as classification technique. Random forest exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5 [6]. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and allowing them to vote for the most popular

class, similar to most classifiers, RF can also suffer from the curse of learning from an extremely imbalanced training data set. Random Forest have proposed ways to deal with the problem of extreme imbalance, based on the Random Forest (RF) algorithm [3] in additions to balanced random forest (BRF) that is weighted random forest (WRF).

WRF incorporate class weights into the RF classifier, thus making it cost sensitive, and it penalizes misclassifying the minority class. But Weighted random forests are computationally less efficient with large imbalanced data since they need to use the entire training set. Also, assigning a weight to the minority class may make the method more vulnerable to noise (mislabeled class) [7]. Balanced Random Forest is a modification of RF, where for each tree two bootstrapped sets of the same size, equal to the size of the minority class, are constructed: one for the minority class, the other for the majority class. Jointly, these two sets constitute the training set [10].BRF combines the sampling technique and ensemble idea of random forest. The other hand, clustering is used to propose new sampling technique procedure in process Balance Random Forest to aim at preventing taking a high number of very similar samples and neglecting some important ones [9].

Clustering is a technique for finding similarity groups in a data, called clusters. Process clustering only to the negative samples. After balanced data, construct the tree in one process with balanced data, construct tree use CART algorithm one of an unpruned decision tree.Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. In this research, because changed the steps of the balanced random forest algorithm, so this purpose method called modified balanced random forest (MBRF).

1.3 Conceptual Framework/Paradighm

There are two variables applied to conduct measurement in this research, namely:

Table 1.1: Variable Conduct Measurement

Variable	**Variable's Information**
Data Splitting	Various data splitting for data testing and data training
RF Parameter	The parameter to induce RF, consisting number of tree induced in the forest.

1.4 Statement Of the problem

Based on the problems discussed in the background, this is the point of problem in research thesis:

1 The customer is one of the main sources of corporate earnings that resulted in customers becoming an important asset to the company. The occurrence customer can lead to decrease in revenue of the company.

2 Customer churn data has an unbalanced data characteristic, where data has one of the sequences with more samples than the other classes.

3 Data mining classification techniques can be used for customer churn prediction. The classification technique can not work well on unbalanced data, since the classification assumes that the data is drawn from the same data distribution, presenting imbalance data to the classifier will produce undesirable results.

4 Balance random forest classification processes balance data approach on random forest classification algorithm, balanced random forest performs combined process of sampling technique (undersampling) and esemble idea. Undersampling has the weakness that the occurrence of loss information when preventing and taking some important ones.

Based on the some of point problems discussed above, the main issue of this study is the process to solve the imbalanced data on predicting customer churn with two main approaches, using the combination of undersampling strategy base on clustering and Balance Random Forest.

1.5 Objective

Based on the statement of the problem and rationale that have been delivered, The objectives of this research are:

1 Create customer churn prediction as Churn Prevention System on Customer churn data in PT Telkom.

2 Handling imbalance data on "Customer Churn Prediction" to improve the effectiveness of Random Forest in producing better prediction performance.

3 We propose to guide the undersampling procedure to improve the effectiveness of Balance Random Forest and provide better performance of balanced random forest.

1.6 Hypothesis

The previous research still presents handling imbalance data in Random Forest used Balance Random Forest. Balance Random Forest has little effect on the classifier produced [7]. The result from the previous research still needs improvement [7] [8]. A combined approach Balance Random Forest and undersampling strategy based on clustering (MBRF) to improve more performance on the classification results and time-consumption.

1.7 Assumption

The global problem in churn prediction includes the variation of the 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 studies are needed to address this problem. This study focuses on the problem of churn prediction accuracy and based on the following assumptions:

1 This research is 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 churn to produce a good performance.

1.8 Scope of work

This section consists of the scopes of knowledge, facility, user, and usability

A Scope of Knowledge

1 Data Mining
Data mining has the meaning of extracting or mining knowledge and large-sized data set. Data mining is also a process of exploration and analysis conducted automatic or semi-automatic in a large set data to find patterns and rules that mean [11].

2 Random Forest
Random Forest is first introduced as one of the classification algorithms in data mining techniques. This algorithm contains a collection of tree decision. Balance Random Forest (BRF) and Weighted Random Forest are developed further Random Forest for use in data imbalance so that the classification process runs well.

B Scope of Facility
Facility is applied in a computer for knowing predict customer churn.

C Scope of User
A User is a person who knows about data mining, Users in this study are used as a student, lecturer, researcher, professional..

D Scope of Usability
Usability is the predicted potential customers who do churn, and as the basis for management and marketing company to take the next strategy. As well as researchers study data mining as a reference .

1.9 Significance of the Study

In this research, the methods could increase the accuracy of the churn prediction and perform under label noise and perform low time consumptions.