CHAPTER 1 THE PROBLEM

Chapter 1 describes the underlying background and the problems of the research, research objective, and the entire conceptual framework of the research. This chapter consists of nine sub-chapters namely: 1) Rationale; 2) Theoretical Framework; 3) Conceptual Framework/Paradigm; 4) Statement of the problem; 5) Objectives; 6) Hypothesis; 7) Assumption; 8) Scope and Delimitation; 9) Importance of the study;

1.1 Rationale

Customer churn has become a significant problem and also a challenge for Telecommunication company. The company is necessary to evaluate the problems that make customer churn; so that the managements can make appropriate strategies to minimize the churn and retaining their customers. In order to survive in a competitive marketplace, telecommunication companies are turning to data mining technique for churn analysis. This approach makes company understand customer's behaviors from their own data so the right CRM (Customer Relationship Management) strategies can be implemented in order to save the revenue.

1.2 Theoretical Framework

Some techniques have been developed to address the imbalanced data and they can be categorized into three groups depending on how they deal with the problems. First, algorithm level approaches may attempt to adapt the existing classifier learning algorithms to bias the learning towards the minority class. Second, data level approaches may rebalance the class distribution by resampling the data space, and the third is cost-sensitive learning framework, it falls between data and algorithm level approaches [1].

As mention above on the data level, the approach tries to decrease the effect caused by imbalance with processing step by changing class distribution. This approach is called sampling technique, there are three categories of sampling; Undersampling methods, which create a subset of the original data-set by eliminating instances (usually majority class instances); Oversampling methods, which create a superset

of the original data-set by replicating some instances or creating new instances from existing one; and the hybrid methods that combine both sampling methods.

Ensemble methods use a set of classifiers to make prediction. The generalization ability of ensemble is usually much stronger than that of individual classifier [1]. Generalization ability indicates how well the unseen data could be predicted by learner trained from the training data. Ensemble methods train a set of a base learner to conduct predictions with each one of them and then combine these predictions to give a final decision by the voting process.

Many approaches to deal with class imbalance problems due to its simplicity and good generalization ability have been developed using bagging ensembles. The hybridization of bagging and data preprocessing techniques are usually simpler than their integration like boosting. A bagging algorithm does not require to recompute any kinds of weight, the key factor is to collect each bootstrap replica, that is how class imbalance problem is dealt to obtain a useful classifier in each iteration without forgetting the importance of diversity.

1.3 Conceptual Framework/Paradigm

The customer of Broadband Internet of PT. Telkom is increasing every day but on the other side there are customers who decide to stop the service. These customers visit Plasa Telkom to register the churning process. This Churn is called quit for “Atas Permintaan Sendiri” (APS). The APS churn data will be used in this research because the amount of this data is very small comparing to the customers so this data churn is imbalanced data.The conceptual framework for churn prediction is described as follows:

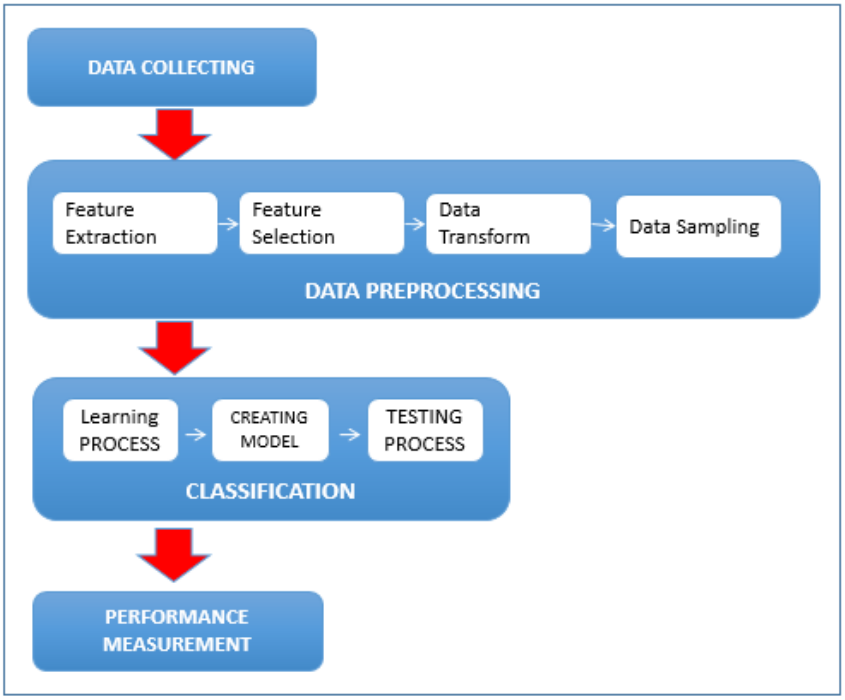

Figure 1-1. Churn Prediction Framework

Dataset broadband internet is collected, then data preprocessing stage is prepared, this process includes feature extraction; feature selection selects the related attribute; then data transform converts the data into numerical type without changing the meaning of the data. The last process is data sampling, it generates a variety dataset using combined SMOTE and RUS.

The next process is classification process, using bagging methods and some classifier to train a set of base learner. From this process, some predictions are resulted then continue to the process voting. This research is expected to handle the imbalance problem and improves the performance of churn prediction.

1.4 Problem Statement

Churn prediction is considered as one of data mining application reflecting the imbalanced problems. Based on data from PT. Telkom Indonesia Regional 6 Kalimantan, the average APS churn rate for Internet broadband customer from July until December 2016 are less than 1% per month and it was very low and categorized as imbalanced data.

Imbalanced dataset problem occurs when one class, usually the one that refers to the concept of interest (positive or minority class), is underrepresented in the data-set and the number of negative (majority) instances outnumbers the number of positive class instances [2]. The lack of data churn is an issue in prediction because it leads to poor performance when classifying minority class and then it may cause the difficulties in developing a good prediction model.

1.5 Objectives

According to the problem's statement above, it leads some objectives to break the problems. Some techniques applied to deal with imbalanced data in order to get the better prediction of churn. Such as sampling, one technique that infers to balance the original data. The most basic sampling methods are random over-sampling and random under-sampling. Random over-sampling duplicates minority class training and random under-sampling eliminate majority class and the objectives of this research are:

1. to handle imbalanced data problems using the combination of sampling methods, they are random over-sampling technique (SMOTE) and Random Under-Sampling (RUS) in order to decrease the degree of class imbalance.

2. to improve the churn prediction performance of single classifier C4.5 by implementing the data preprocessing and bagging, one of ensemble methods in classification. Hypothesis

Hypothesis in this research is the combination of SMOTE and RUS in sampling can handle the imbalanced data and can obtain the increase value of F-Score when applying bagging methods in classification process.

1.6 Assumption

The assumptions of this study are:

1. This study is conducted based on voluntary and deliberates data churn, it was APS data churn of Indihome product in Telkom Regional 6 Kalimantan.
2. This study discusses issues how to handle the imbalanced data churn for prediction.
3. This study allows some scenario experiments to build churn prediction model that has a good performance.

1.7 Scope and Delimitation

Scope and delimitation of this research are as follows :

1. This study attempts to compare the performance of a churn prediction using single classifier without data preprocessing, and the combination of SMOTE and RUS with Bagging methods.
2. This study uses C4.5 as single classifier.
3. The dataset used in this research is churn APS of Broadbnad Internet from Telkom Regional 6 Kalimantan.

1.8 Importance of the Study

This research wants to evaluate the performance of combined SMOTE with RUS for rebalancing the distribution of an extremely imbalanced data. Then by using bagging methods of a single classifier C4.5 in classification process will improve the performance of churn prediction model and it can give a positive contribution in handling churn's problem in PT. Telkom specially Telkom Regional 6 Kalimantan.