CHAPTER 1
INTRODUCTION

In the world which the communication and information technology is rapidly developed and digital information across a wide variety of fields becomes more abundant and accessible, new computational theories and practical tools to extract and select useful information become critical and urgently required. These notions are the subject of knowledge discovery in database (KDD) [1]. KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately human understandable patterns in data. It concerns with the development of methods and techniques for analyzing data in various forms according to the users' objective. It plays an essential role in various applications from daily life activities to worldwide business. Each step of the KDD process is displayed in figure 1.

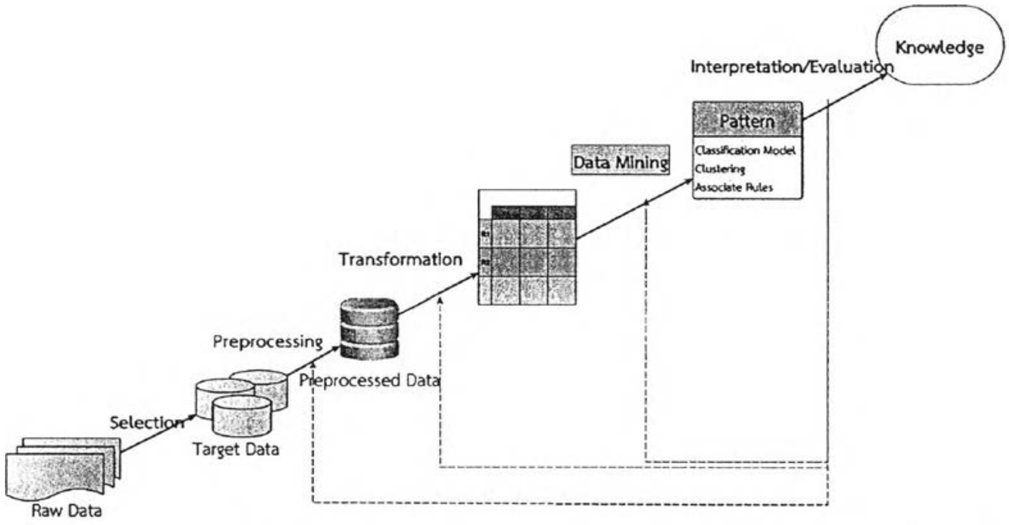

Figure 1: Knowledge discovery in databases process (KDD process)

Data mining is an important part in the KDD process because it transforms a dataset into useful knowledge. According to SAS [2], the data mining process composes of sampling, exploring, modifying, modeling and assessing large amount of data to uncover a previously unknown pattern. It combines applications and concepts from applied mathematics, statistics and artificial intelligence to deal with big data. Data mining is widely used in various fields such as business, scientific researches, engineering, medicine etc. In business [3], data mining could be used to determine relationships among internal factors such as price, marketing strategy for products, staff contribution and external factors such as competitions from rival companies, customer responses, etc. These relationships can give useful summarized information from the detailed transaction data and help making decisive business decisions in time. In scientific researches [4], it can be applied to help scientists to analyze their raw experimental results and provide guidelines, patterns or relationship that they can further explore. In the field of engineering [4], data mining can be applied in the process of modeling and optimization, machine diagnostics and predictive maintenance.

In medicine [5], there is a need for an efficient analytic methodology for detecting valuable information from a large number of medical records. Data mining can be applied on various medical processes such as a diagnosis process. Using data mining techniques through historical records, it can help physicians and medical personals to select suitable and available treatments which fit the requirement of their patients. It also predicts the outcome of the treatment and suggests the idea to improve the result of existing ones. Many domains of bioinformatics researches [6] also apply data mining for the knowledge extraction such as genomics, proteomics microarrays, and system biology since there are large-sized datasets that are needed to be handled in each domain.

There are various tasks in data mining, each of which has different requirements and usages. Classification is a data mining task which concentrates on finding a relationship, rule or pattern from a multi-attribute dataset for predicting the pre-defined target class of unknown instances. The relationship or pattern gathered from classification techniques is generalized from the dataset. This pattern is further utilized to predict a class of an unknown instance from the same domain. Examples of models in classification are decision tree [7], naïve Bayes classifier [8], neural network [9], support vector machine [10], *k*-nearest neighbor [11]. Another task in data mining is clustering which focuses on finding the similarity and dissimilarity of a group of instances inside a dataset and grouping instances based on their similarity. The similarity for clustering could be the distance between instances, the density around instances, etc. Examples of clustering techniques are *k*-mean clustering [12], hierarchical clustering [13], DBSCAN [14], etc. Regression [15] is also categorized as a task in data mining which is used to find a function that explains the correlation among instances. The function can be linear or non-linear depending on characteristics of a dataset. Association rule induction [16] is another task which focuses on extracting frequent patterns or relationships among a set of data items in the repository. Examples of association rules techniques are apriori algorithm [17], FPtree [18]. Among these data mining tasks, this dissertation concentrate on a special problem in the classification task called "Class imbalance problem" [19].

1.1 Introduction to class Imbalance problem

Classification techniques are developed to deal with various kinds of datasets whose target class is defined. With a sufficient amount of instances for each class, many classifier algorithms provide the classification model that can effectively represent the existing dataset and predict the class of unknown instances. However, in real-life problems, there are many datasets which have imbalanced distribution of classes. One or some classes may have the number of instances relatively less than others. Some standard classification algorithms such as a decision tree induction or a support vector machine are biased toward the majority class and treat the minority class as noises. Since most classifiers aim to maximize the overall accuracy, instances from that minority class might be misclassified in order to maximize a high prediction rate of instances in the majority class. This behavior is reasonable with respect to accuracy of the whole dataset. But for some problems such as the detection of oil spills in satellite radar images [20], the detection of fraudulent telephone calls [21], in-flight helicopter gearbox fault monitoring [22], information retrieval and filtering [23] and diagnosis of rare medical conditions such as thyroid diseases [24], the prediction on the class with fewer instances becomes the main focus of the classification. For example, in the diagnosis of rare medical conditions, doctors want to identify a pattern of a rare medical condition from a large-sized medical record. The classification problem dealing with this kind of dataset is called **class imbalance problem** and it is the central theme of this dissertation.

Despite the fact that a multiclass classification problem is more frequently found in real-life problems, it is much simpler to deal with two-class cases and extends it to cover the multiclass ones. So, only the binary classification problem is focused in this dissertation. For the binary classification, the class with a smaller number of instances is defined as minority or positive, while the other class with a larger number of instances is assigned as majority or negative. This definition is used for the entire dissertation.

There are many approaches to deal with class imbalance problem which are described later in chapter 2. Among all approaches, oversampling techniques which are used on imbalanced datasets during the data-preprocessing stage is emphasized in this dissertation since the resulting balanced dataset after applying these techniques is applicable for any classifiers. Synthetic Minority Oversampling

Technique (SMOTE) by Chawla [25] is one of prominent oversampling techniques widely used for class imbalance problem. In SMOTE, a new synthetic instance is generated between each positive instance and its nearby positive instance. A collection of synthetic instances generated by this idea has effectively improved the performance on predicting positive instances. With its simplicity, there are many researches applying the idea from SMOTE such as ADASYN [26], borderline-SMOTE [27], safe-level SMOTE [28] and DBSMOTE [29] in order to further improve the classification performance. However, there are some flaws of SMOTE and its successors which should be explored such as a problem with some positive instances which are located far from groups of positive instances, a problem with synthetic instances which are located near existing negative instances and the selection of the parameter *k* in SMOTE. Therefore, this dissertation works on developing the framework for oversampling techniques in order to deal with these issues in a binary class imbalance problem more effectively. The framework is introduced by providing some new oversampling techniques that can remedy these weaknesses. These new oversampling techniques are expected to have a better accuracy performance based on some performance measures.

Classifiers used in this dissertation and existing oversampling techniques are introduced in chapter 2. New oversampling techniques in the framework are described in chapter 3. The result and improvements caused by these new techniques are presented in chapter 4. The discussion and conclusion are drawn in chapter 5.

1.2 Objectives

The objectives of this dissertation are to:

1. provide a method to improve the performance of some existing minority oversampling techniques
2. introduce a minority oversampling technique which uses a different approach from SMOTE and provides a statistically significantly better performance

To fulfil these objectives, one additional process called minority outcast handling with 1-NN and two new oversampling techniques are introduced. Minority outcast handling with 1-NN is the approach to handle some instances called minority outcast instances in order to improve the classification performance. It is included in both new oversampling techniques that are introduced in this dissertation. The first oversampling technique introduced in this dissertation is relocating safe-level SMOTE (RSLS) which is a modified version of safe-level SMOTE [28] to reduce the number of conflicting synthetic instances surrounding by negative instances. The other oversampling technique is adaptive neighbor SMOTE (ANS) which provides the procedure of automatically selecting the parameter k of SMOTE [25] for each positive instance instead of using one static value for every positive instance. The classification performances from these two oversampling techniques are shown and compared against ones from other oversampling techniques. The significance of result improvements caused by these two techniques is tested with Wilcoxon signed-rank test [30].

1.3 Scope of Work

This framework concentrates on the binary class imbalanced dataset. All datasets are complete with no missing values. Every attribute in each imbalanced dataset used for this work is continuous since the distance calculation and the creation of synthetic instances are performed with the Euclidian distance. In the experiment, oversampling techniques introduced in this framework are compared with various oversampling techniques. The performance is evaluated with F-measure, geometric mean and adjusted g-mean [31]. Five classifiers used in this dissertation are decision tree (C4.5) [32], naïve Bayes classifier [8], multilayer perceptron [9], support vector machine [10] and k-nearest neighbor [11]. All oversampling procedures are performed in the R programming environment [33] and the classification and evaluation processes are performed in KNIME [34].

1.4 Expected Outcome

In 70 cases from 14 imbalanced datasets and 5 classifiers, the result from suggested oversampling techniques in this framework are expected to achieve the best or top three values on F-measure, geometric mean and adjust g-mean. The difference of performance measure values from suggested oversampling techniques against the others are expected to be significantly positive for all classifiers after testing with Wilcoxon signed-rank test.