CHAPTER 1

INTRODUCTION

The digital revolution that is witnessed these days has resulted in an exponential availability of information and data on the Internet and in digital database systems. Machine Learning (ML) algorithms are widely used to automate the process and extract the knowledge from those data. Knowledge Data Discovery (KDD) is the process of extracting or mining knowledge from large amounts of data (Tipawan 2012). But the quality of discovered knowledge is highly depends on data quality. Unfortunately such contemporary data available may be prone to noise, incompleteness, redundancy and inconsistency. This kind of data can be the reason uncertainty to the process of mining, by producing unreliable output. Accuracy, comprehensiveness, reliability, computational time, interpretability and accessibility are the basis on which data is preprocessed to extract high quality knowledge. D*ata cleaning, data integration, data transformation and data reduction* are the preprocessing techniques (Han 2006).

The main task of *data cleaning* is filtering or removing noises by dealing with outliers, filling-in unknown values and inconsistencies resolution (Gonzalo et al. 2010). Missing values can be handled by ignoring the instances completely or by filling empty positions with values manually inserted or using a constant which can be equal to each instance or by computing mean or median or by using other attributes to predict values. Outliers can be handled by using binning, clustering, regression and manual inspection. Inconsistent data are mainly due to the data-entry error or poor data integration. They can be rectified by having data-entry verification and external reference.

Data cubes or warehouses are formed by merging data accessed from several sources in the process of *Data integration*. *Data transformation* (Tipawan 2012) by smoothing, aggregation (summarization, forming data cube), generalization (replace data with higher level concepts) and normalization (scaling to within a specified range) improve the performance of mining algorithms on the basis of accuracy and efficiency.

Data warehouses may contain massive data influencing adversely increase in running time of mining algorithms. The use of algorithms seems to be ineffective for high dimensionality. Hence, data cube aggregation, dimensionality reduction by eliminating redundant features (Feature Selection Technique), data compression, numerosity reduction and data discretization can be used as *data reduction* techniques. KDD offers a global framework to prepare data in the right form to extract high quality knowledge. Thus, the preprocessing step is an important step resulting in Data Mining operations.

1.1 NEED OF THE RESEARCH WORK

In this competitive world, Data Mining faces the growing challenge of systematic extraction of knowledge in large datasets, as the real world data are generally incomplete, inconsistent, redundant and noisy. Faulty data collections, data entry problems, limitations in technology and transmission issues may contribute to noisy data. It is being tried to eliminate incomplete, inconsistent and noisy data by preprocessing the data before the learning process begins. But the result of preprocessed data still leads to increased misclassification rate during the learning process. So, the main aim of this research is to improve the learning process (classifier) performance through the implementation of preprocessing techniques as well as assess the performance of the classifier by using Receiver Operating Characteristics (ROC) curve measure.

1.2 MOTIVATION

Extremely high rate of noises, high dimensionality and inconsistency among the data pose great issues and challenges for the successful application of learning process such as classification, clustering and association rule mining to applications like fraud detection, medical health care, customer segmentation etc. It also poses great issues and challenges for the successful application of preprocessing techniques such as discretization, missing value imputation, feature selection and data reduction.

- **Issues considered for discretization:**
    - *Increased computational time while finding the midpoint of all the adjusting pairs of data for boundary point selection during discretization of continuous attribute/feature.*

  - *During discretization, large number of intervals are generated which leads to increased size of the tree while mining.*
  - *Interdependency between the class feature and the other features values, which have an impact on the accuracy performance of the classifier.*
- **Challenges for discretization:** *Reduced computational time, increased interdependency and generation of less number of intervals.*

Traditional discretization methods are not able to work well as the class features are not taken into account while finding the boundary points for the intervals and the initial boundary points are taken by finding the midpoints of all the two adjacent data points. These issues motivated to find a solution to meet the challenges.

- **Issue considered for missing value imputation:**
  - *The unknown values in the dataset give the problem of high misclassification rate.*
- **Challenge for missing value imputation:**
  - *Increased accuracy performance.*

Traditional ways of eliminating all instances (values) with missing values are considered inappropriate and produces more biased estimates. It is hazardous due to loss of sample size. The drawbacks of mean imputation by using entire data or sampled data are overestimation of sample size, underestimation of variance and negatively biased correlation. Imputation using distance measures is a time consuming process and imputation using the observed values taken randomly and using most frequent values degrades the accuracy of the classifier. These issues motivated to develop a solution for filling the missing values.

- **Issues considered for feature selection:**
    - *Many real world data contains historic and irrelevant information. Irrelevant information degrades the performance of the classifier both in computational time (due to large set of feature set) and the estimated accuracy (due to uncorrelated information).*
    - *Applying feature selection only on the sample of data yields a worst performance of the classifier as the data is growing rapidly at faster rate.*
- **Challenges for feature selection:** *Increased accuracy performance and reduced computational time.*

Traditional feature selection technique RELIEF algorithm does not deal with the noisy and incomplete datasets and random sampling being used for selecting the sample of values. This method will lead to time consuming. Also, sample size must be given by the user as input parameter. These are the striking problems of the state-of-the-art preprocessing methods. These issues motivated to develop a solution for selecting the relevant features/attributes.

In medical health care process, the Government and the public needs the medical community to give explanation on possible threat of diagnostic procedures. So, diagnosing either positive or negative results of cancer disease for a patient is serious issues nowadays by using the classifier models in Machine Learning scenario. Any assessment of diagnostic performance seems to be required before taking decisions.

- **Issues considered for classification performance:**
    - *Lack of classifier performance in medical health care due to its inconsistency and its inability to discriminate between classes.*
- **Challenges for classification performance :** *Assessing the classification performance*

1.3 CONTRIBUTIONS

In this thesis, a systematic study is made to solve the existing problems of the recent preprocessing techniques. The contributions are described in detail below.

- NAD (*Normalized And Discretization method*) algorithm is proposed to handle the problems of computational time while discretizing the continuous features/attributes. The contribution is, in the proposed method, the boundary points are selected by finding the midpoints where the two consecutive values have different class values instead of choosing the midpoint between the two consecutive pair of values. The number of intervals at the initial stage itself is reduced. Since the instances belonging to the same classes are considered together, interdependency between the class and the continuous features are maximized.

- DISCEND algorithm proposed for imputing a value in the place of unknown values. The unknown values are filled in by the lowest mean value in each interval during the initial stage of discretization process. This approach increases the accuracy of the classifier (that is it reduces the misclassification rate) and the relevancy between the instances and attributes.

- RELIEF-DISC algorithm is proposed to handle the issues described earlier for selecting the features/attributes. Selecting the instance in done by random sampling in RELIEF. In RELIEF-DISC, the instance is chosen from each interval while doing discretization instead of applying random sampling. This will increase the speed of retrieving by preserving the quality of features.

- Extended RELIEF-DISC concept handles the noisy and incomplete data as RELIEF-DISC fails to do so. Integrating all the preprocessing techniques increase the overall performance (accuracy) of the classifier and reduce the computational time. First, the missing values are filled during the initial stage of discretization. Then the discretization is completed and the relevant features are selected. The resultant preprocessed data is applied in the classifiers to show the improvement in the accuracy performance.

- ROC curve assessment for evaluating the performance of C4.5 decision tree classifier and Naïve Bayes classifier using the existing preprocessing techniques and the proposed methods.

1.4 THESIS STATEMENT

Real world raw data are supposed to be noisy, redundant and inconsistent. Preprocessing can be done by many techniques. This research tried to preprocess the continuous attributes during Knowledge Data Discovery before applying learning/mining process, as many applications contains only the continuous attributes. Filling the missing values, discretizing the filled in data and selecting the features which are relevant to the mining process are the three steps in preprocessing of raw data. These three preprocessing steps are integrated and implemented for the improvement of the performance of the classifiers and for the reduced computational time.

In the first step, the missing values in the raw data are filled in while discretizing the continuous features in the initial step of finding the boundary points. After the missing values are filled in, discretization process is completed in the second step. Once all the continuous features/attributes are discretized, the next step in the integrated process is to choose only the relevant features for doing the mining procedure. The integrated implemented process increases the accuracy performance of the C4.5 classifier and Naïve Bayes Classifier ie. it reduces the misclassification rate with the reduced computational time. Also an assessment is made on the performance of the classifiers by using ROC curve assessment method.

Experimental comparison is made using tables, graphs and curves to show the improved performance than the existing methods. The entire proposed framework is depicted in Figure 1.5.1.

1.5 PROPOSED FRAME WORK

The frame work given in Figure 1.5.1 depicts the procedure of discretizing the data taken from the dataset and the unknown values are filled in. The discretized data is stored in the database which is retrieved by the feature selection method to choose the related features. The preprocessed data are classified by the classifiers and the result is assessed by the ROC curve assessment process.

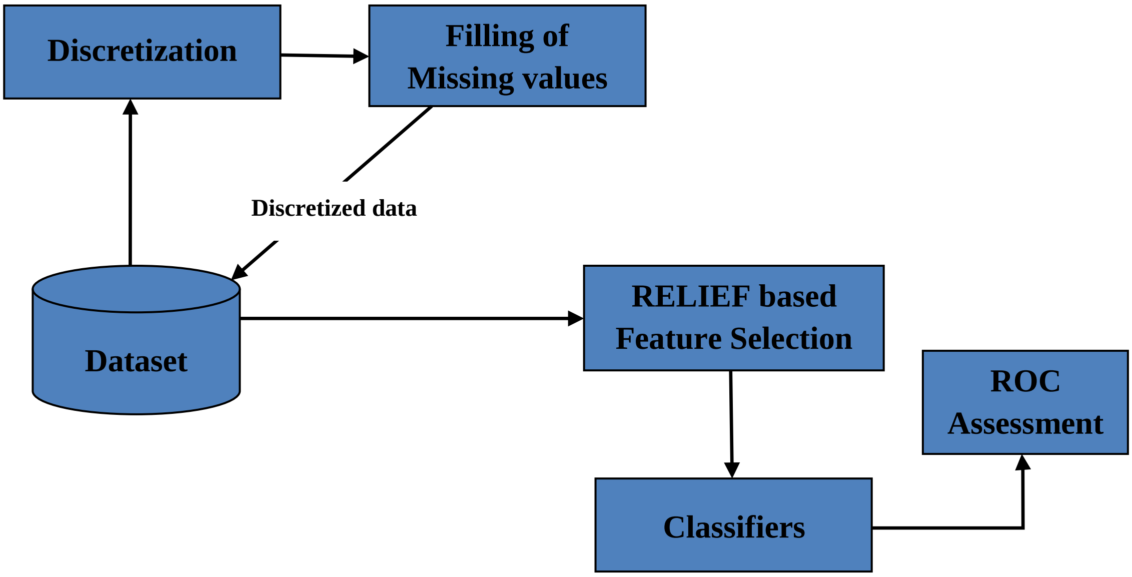

Figure 1.5.1 : Proposed Frame Work

1.6 THESIS OVERVIEW

The structure of remaining thesis is as follows:

Chapter 2 illustrates the Knowledge Discovery in Databases process. It provides a formal definition for it and considers in more details the preprocessing steps in KDD and the Data Mining phases of the process. Chapter 3 reviews the works related to the research work. A brief review on discretization, feature selection and missing value imputation is done. Also, ROC curve assessment is briefly studied. Chapter 4 introduces the concept of NAD. It presents the methodology and the algorithm developed to exploit prior knowledge in driving the discretization process. Chapter 5 gives a brief description of DISCEND algorithm for imputing the missing values. The methodology and the algorithm are explained in detail. Chapter 6 describes the concept proposed on feature selection method. RELIEF-DISC methodology and the algorithm developed are explained in this chapter. Chapter 7 describes the concept of integrated method implemented. The entire process is explained and discussed in this chapter. Chapter 8 shows the ROC curve assessment of the C4.5 classifier and Naïve Bayes classifier by using existing preprocessing techniques and the proposed method. Chapter 9 draws some conclusions and the future work.