CHAPTER 1

INTRODUCTION

1.1. DATA MINING IN LARGE DATABASE ANALYSIS

Data mining is an interdisciplinary subfield of computer science used in various applications of engineering and technology as discussed by Hastie et al (2001). The process involves a series of computations of large data sets wherein patterns are discovered that incorporate techniques at an intersectional stage wherein statistics, artificial intelligence, database systems and machine learning concur. Basic benefit from the process of data mining process from any data set is the extraction of information and then its subsequent transformation into a simplified and easy to relate structure for complex purposes. Besides raw analysis step, the process involves database as well as data management facets, model & inference and complexity considerations, data pre-processing, post-processing of discovered structures, visualization as well as online updating.

The commercial success of database technology and the availability of relatively inexpensive sensing, storage, and processing hardware have led to explosive growth in online data storage over the last two decades. In turn, these large databases have motivated the rapid development of data mining and knowledge discovery, namely, the search for structure in large volumes of data.

While science and industry have scaled up their data gathering activities, traditional data analysis research in statistics and machine learning has been relatively slow to take up the challenge and much research and published work is still focused on relatively small data sets.

It is clear that in the near future, large data sets will eventually play an essential role in data-analytic research settings.

The most fundamental challenge for the large data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real-time because storing of all observed data is nearly infeasible.

Real-world data are dirty, and therefore, noise handling is a defining characteristic for data mining research and applications. A typical data mining application consists of four major steps: data collection and preparation, data transformation and quality enhancement, pattern discovery, and interpretation and evaluation of patterns . It is expected that the whole process starts with raw data and finishes with the extracted knowledge. Because of its data-driven nature, previous research efforts have concluded that data mining results crucially rely on the quality of the underlying data, and for most of the data mining applications, the process of data collection, data preparation, and data enhancement cost the majority of the project budget and also the developing time circle. However, data imperfections, such as erroneous or inaccurate attribute values, still commonly exist in practice, where data often carry a significant amount of errors, which will have negative impact on the mining algorithms.

Data mining, or knowledge discovery, has become an indispensable technology for businesses and researchers in many fields. Rapid advances in

data collection and storage technology have enabled organizations to accumulate vast amounts of data. However, extracting useful information has proved extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small.

Large datasets are usually complex in structure and challenging in extracting meaningful information from them. These major problem can be solved in two ways through datamining Techniques. One way of extracting information is through information measures of computational and algorithmic complexity; another is through quantifying variability in these sets.

The goal of this archive is to store large data sets that span a wide variety of data types and problem tasks. Understanding the relevant factors in the analysis of large datasets is an important step in improving study design, structured analysis of data, and generalizability of results. Thus, analysis and investigation of large datasets purely depend on the data mining techniques. The present research work focuses on utilizing the appropriate data mining techniques for analyzing the large datasets.

1.2. KNOWLEDGE DISCOVERY IN DATABASE

The transformation of data into knowledge has been using mostly manual methods for data analysis and interpretation, which makes the process of pattern extraction of databases too expensive, slow and highly subjective, as well as unthinkable if the volume of data is huge. The interest in automating the analysis process of great volumes of data has been fomenting several research projects in an emergent field called Knowledge Discovery in Databases (KDD). KDD is the process of knowledge extraction from great

masses of data with the goal of obtaining meaning and consequently understanding of the data, as well as to acquire new knowledge. This process is very complex because it consists of a technology composed of a group of mathematical and technical models of software that are used to find patterns and regularities in the data.

KDD has been defined as the process of discovering valid, novel, and potentially useful patterns from data. An important notion, called interestingness, is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity.

To extract knowledge from databases, it is essential that the Expert follows some steps or basic stages in order to find a path from the raw data to the desired knowledge. The KDD process organizes these stages in a sequential and iterative form. In this way, it would be interesting if the obtained results of these steps were analyzed in a more interactive and friendly way, seeking a better evaluation of these results. The process of knowledge extraction from databases combines methods and statistical tools, machine learning and databases to find a mathematical and/or logical description, which can be eventually complex, of patterns and regularities in data. The knowledge extraction from a large amount of data should be seen as an interactive and iterative process, and not as a system of automatic analysis.

The interactivity of the KDD process refers to the greater understanding, on the part of the users of the process, of the application domain. This understanding involves the selection of a representative data subset, appropriate pattern classes and good approaches to evaluating the knowledge.

KDD Process: Knowledge discovery from data can be understood as a process that contains, at least, the steps of application domain understanding,

selection and preprocessing of data, Data Mining, knowledge evaluation and consolidation and use of the knowledge. The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user.

The KDD process begins with the understanding of the application domain, considering aspects such as the objectives of the application and the data sources. Next, a representative sample (e.g. using statistical techniques) is removed from database, preprocessed and submitted to the methods and tools of the Data Mining stage with the objective of finding patterns/models (knowledge) in the data. This knowledge is then evaluated as to its quality and/or usefulness, so that it can be used to support a decision-making process.

1.2.1. Challenges in Data Analysis

Traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by new data sets. The following are some of the specific challenges that motivated the development of data mining:

Scalability: Because of advances in data generation and collection datasets with sizes of gigabytes, terabytes, or even megabytes are becoming common. If data mining algorithms are to handle these massive datasets, then they must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. For instance, out-of-core algorithms may be necessary when processing data sets that cannot be fitted into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

High Dimensionality: It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. If the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken.

Traditional data analysis techniques that were developed for low-dimensional data often did not work well for such high-dimensional data. Also, for some data analysis algorithms, the computational complexity increases rapidly as the dimensionality (the number of features) increases.

Heterogeneous and Complex Data: Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. Recent years have also witnessed the emergence of more complex data objects. Examples of such non-traditional types of data include collections of Web pages containing semi-structured text and hyper lines; DNA data with sequential and three-dimensional structure; and climate data that consists of time series measurements (temperature, pressure, etc.) at various locations on the Earth's surface.

Techniques developed for mining such complex objects should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity, and parent-child relationships between the elements in semi-structures text and XML documents.

Data Ownership and Distribution: Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques. Among the key challenges that faced distributed data mining algorithms include (1) how to reduce the amount of communication needed to perform the distributed computation, (2) how to effectively consolidate the data mining results obtained from multiple sources, and (3) how to address data security issues.

Non-Traditional Analysis: The traditional statistical approach is based on a hypothesize- the test paradigm. In other words, when a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis. Unfortunately, this process is extremely labor-intensive. Current data analysis tasks often require the generation and evolution of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Furthermore, the data sets analyzed in data mining are typically not the result of a carefully designed experiments but often represent opportunistic samples of the data, rather than random samples. Also, the data sets frequently involve non- traditional types of data and data distributions.

The present research work mainly focuses on handling the above said challenges through three steps namely preprocessing, feature selection and classification.

1.3. GENERAL OVERVIEW OF PREPROCESSING

Data preprocessing is a broad area and it consists of a number of different strategies and techniques that are interrelated in complex ways. We

will present some of the most important ideas and approaches, and try to point the interrelationships among them. The preprocessing techniques fall into two categories: selecting data objects and attributes for the analysis or creating/ changing the attributes. In both cases the goal is to improve the data mining analysis with respect to time, cost, and quality. Specifically, following are the important preprocessing techniques:

Prior to using data mining algorithms, assembling target data set is imperative. Data mining basically can uncover existing patterns in the data, as well as target data set has to be large to contain existing patterns while maintaining it brevity enough so that it may be mined in line with the acceptable time limit. Data marts or warehouses are in general frequently deployed sources of data. Pre-processing is important for multivariate data sets analysis prior to data mining. Thereafter which target set is cleaned. Data cleaning eliminated those observations that have noise and possess missing data.

1.3.1 Types of Preprocessing Techniques

Susceptibility factors with respect to raw data are high especially in relation to noise, missing values, as well as inconsistency. Data mining results are affected by data quality. Therefore to improvise data quality and furthermore mining results preprocessing of raw data needs to be carried out to improvize mining technique efficiency and ease. Data preprocessing as such is perhaps the most significant as part of the data mining process which incorporates initial dataset preparation and transformation. As such if there is an excess of irrelevant and redundant information or there exists unnecessary noise as well as unreliable data, then the process of knowledge discovery in training phase becomes tough. There is a lot of time that can be taken up by steps of data preparation and filtering. Data pre-processing comprises of cleaning, selection, transformation, normalization, and feature extraction and

others. Outcome from data pre-processing is then considered as the final training set. Data preprocessing techniques may be segregated and categorized as follows:

- Data Cleaningcan be applied to remove noise and correct inconsistencies in the data.
- Data Integrationmerges data from multiple sources into a coherent data store, such as a data warehouse.
- Data Transformation, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.
- Data Reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These techniques are not mutually exclusive; they may work together.

Various preprocessing techniques exist, such as:

1. Data cube aggregation: This aggregation basically refers to those operations applicable to data in the construction of a data cube.
2. Dimension reduction: this is where detection of dimensions or attributes that are irrelevant, inadequately relevant or basically redundant. In the following research work dimensionality reduction is carried out using PCA, KPCA. Principal Component Analysis or PCA is basically one that is applicable widely in computer science. Kernel Principal Component Analysis or KPCA is basically one that is an improved PCA, that pulls out principal components through the deployment of

the nonlinear kernel method as presented by Chen et al (2008), Liao et al (2008) and Ding et al (2009).

3. Data compression: where encoding mechanisms are used to reduce the data set size.

4. Numerosity reduction: where the data are replaced or estimated by alternative, smaller data representations such as parametric models, or nonparametric methods such as clustering, sampling, and the use of histogram.

5. Discretization and concept hierarchy generation: where raw data values for attributes are replaced by ranges or higher conceptual levels.

The present research work mainly concentrates on the data cleaning preprocessing technique to fill in the missing values. The major techniques involved in data cleaning process are

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Research and study presented here emphasizes on missing values attributes that possess the nearest neighbor methods. Analysis of data using data mining techniques may be incomplete noisy. Incomplete, noisy, and inconsistent data here generally refers to the commonplace properties of large, real-world databases and data warehouses. Incomplete data can may be inherent on account of various factors. Attributes of interest are not likely to available always as such, like customer information utilized for sales transaction data. Data available otherwise may not be necessarily be included simply as it may not be termed as important at entry point. Relevant data also

necessarily may not have been recorded on account of misunderstanding, or existing equipment malfunctions.

Missing Values: If it is noted that there are many tuples that have no recorded value for several attributes, then the missing values can be filled in for the attribute by various methods described below:

- Ignore the tuple: This is usually done when the class label is missing. This method is not very effective, as the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.
- Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like “Unknown", or -∞. If missing values are replaced by, say, “Unknown", then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common that of “Unknown". Hence, although this method is simple, it is not recommended.
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class as the given tuple.
- Use the most probable value to fill in the missing value: This may be determined with inference-based tools using a Bayesian formalism or decision tree induction.

Missing data randomness can be divided into three classes, as proposed by Little et al (1987):

- Missing Completely At Random (MCAR). This is the highest level of randomness. It occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data. In this level of randomness, any missing data treatment method can be applied without risk of introducing bias on the data;
- Missing At Random (MAR). When the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of the missing data itself;
- Not Missing At Random (NMAR). When the probability of an instance having a missing value for an attribute could depend on the value of that attribute.

There are several methods for treating missing data available in the literature. Many of these methods, such as case substitution, were developed for dealing with missing data in sample surveys, and have some drawbacks when applied to the Data Mining context. Other methods, such as replacement of missing values by the attribute mean or mode, are very naive and should be carefully used to avoid insertion of bias. In a general way, missing data treatment methods can be divided into three categories:

- Ignoring and discarding data. There are two main ways to discard data with missing values. The first one is known as complete case analysis; it is available in all statistical programs and is the default method in many programs. This method consists of discarding all instances (cases) with missing data.

- Parameter estimation. Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm (Dempster 1977) can handle parameter estimation in the presence of missing data;

- Imputation. Imputation is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values.

1.3.2. Imputation Methods

Imputation methods involve replacing missing values with estimated ones based on some information available in the data set. There are many options varying from naïve methods like mean imputation to some more robust methods based on relationships among attributes.

1.3.3. Importance of Preprocessing Techniques

Data in the real world is dirty incomplete with lacking attribute values, lacking certain attributes of interest, or containing only aggregate data ,noisy data and inconsistent. Moreover, in most of the real world datasets, data is not always available. For example, many tuples have no recorded value for several attributes, such as customer income in sales data. So, in order to find the data availability and incompleteness attribute information, data preprocessing methods are essential.

Data preprocessing techniques improve the data quality results and reduce the dimensionality of data with larger dataset analysis. Less quality of data in the decision, duplicate or missing data may cause incorrect or even

misleading statistics. Data preprocessing methods fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

1.4 GENERAL OVERVIEW OF FEATURE SELECTION

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not.

Motivation for applying Feature Selection (FS) techniques in bioinformatics has shifted from being an illustrative example to becoming a real prerequisite for model building. In particular, the high dimensional nature of many modeling tasks in bioinformatics, going from sequence analysis over microarray analysis to spectral analyses and literature mining has given rise to a wealth of feature selection techniques being presented in the field. In this review, focus on the application of feature selection techniques is made. In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. using information theory), feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables, hence offering the advantage of interpretability by a domain expert.

In real-world data, the representation of data often uses too many features, but only a few of them may be related to the target concept. There

may be redundancy, where certain features are correlated so that is not necessary to include all of them in modelling; and interdependence, where two or more features between them convey important information that is obscure if any of them is included on its own.

Generally, features are characterized as follows:

- Relevant: These are features which have an influence on the output and their role cannot be assumed by the rest
- Irrelevant: Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for each example.
- Redundant: A redundancy exists whenever a feature can take the role of another (perhaps the simplest way to model redundancy).

While feature selection can be applied to both supervised and unsupervised learning, the present research work focuses on the problem of supervised learning (classification), where the class labels are known beforehand as discussed by Zheng et al (2008).

Feature selection algorithms in general have two components: a selection algorithm that generates proposed subsets of features and attempts to find an optimal subset; and an evaluation algorithm that determines how good ‘a proposed feature subset is, returning some measure of goodness to the selection algorithm. However, without a suitable stopping criterion the feature selection process may run exhaustively or forever through the space of subsets.

1.4.1 Types of Feature Selection

Ideally, feature selection methods search through the subsets of features, and try to find the best one among all the competing candidate

subsets according to some evaluation function. However, this procedure is exhaustive as it tries to find only the best one. It may be too costly and practically prohibitive, even for a medium-sized feature set size. Other methods based on heuristic or random search methods; attempt to reduce computational complexity by compromising performance.

In Langley (1994) different feature selection methods are grouped into two broad groups (i.e., filter and wrapper), based on their dependence on the inductive algorithm that will finally use the selected subset. Filter methods are independent of the inductive algorithm, whereas wrapper methods use the inductive algorithm as the evaluation function. Wrapper methods wrap the feature selection around the induction algorithm to be used, using cross-validation to predict the benefits of adding or removing a feature from the feature subset used

A strong argument for wrapper methods is that the estimated correlation coefficient of the learning algorithm is the best available heuristic for measuring the values of features. Different learning algorithms may perform better with different feature sets, even if they are using the same training set (Blum & Langley 1997). The wrapper selection methods are able to improve performance of a given regression model, but they are expensive in terms of the computational effort. The existing filter algorithms are computationally cheaper, but they fail to identify and remove all redundant features. In addition, there is a danger that the features selected by a filter method can decrease the correlation coefficient of a learning algorithm.

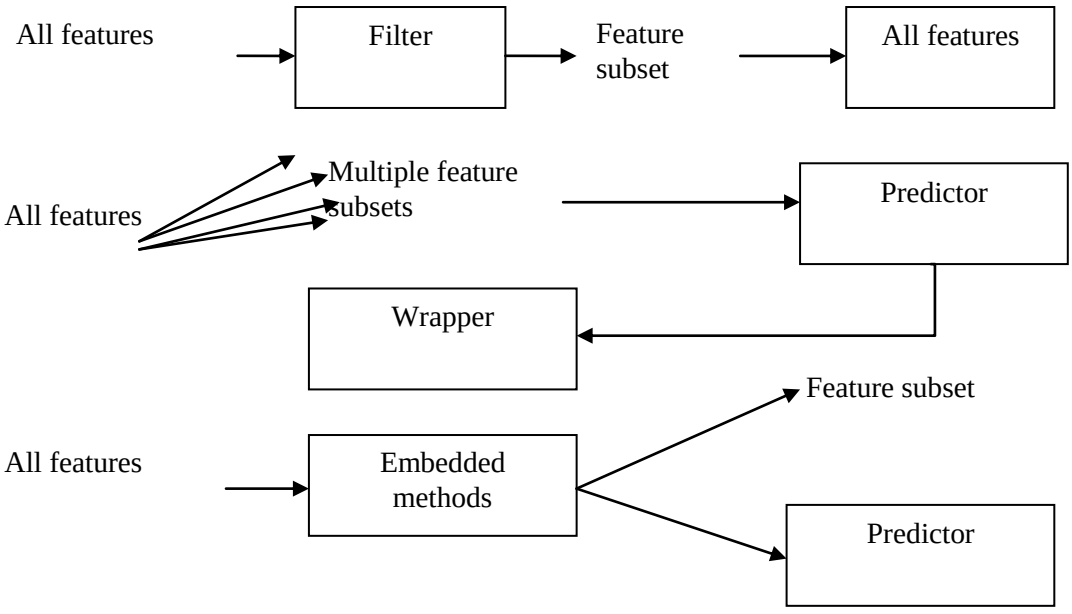

Figure 1.1 Feature selection with strategies

1.4.2 Issues in Feature Selection

Many strategies have been exploited for the task of feature selection, in an effort to identify more compact and better quality feature subsets. Such techniques typically involve the use of an individual feature significance evaluation, or a measurement of feature subset consistency, that work together with a search algorithm in order to determine a quality subset.

It works on the basis of estimating the number of times a certain feature would have occurred in a dataset if the dataset was perfectly representative of the problem domain. Recent trends in developing feature selection methods focus on evaluating a given feature subset as a whole instead of measuring on an individual feature basis.

Feature selection has become the essential step in many data mining applications. Using a single feature subset selection method may generate local optima. Ensembles of feature selection methods attempt to combine multiple feature selection methods instead of using a single on.

The present research work utilizes both hybrid feature selection technique and ensemble feature selection techniques for improving the ovrearall peformance of the classification system. Hybrid feature selection is been widely used in various applications as discussed by Simona et al (2010).

1.4.3 Ensemble Feature Selection Approach

The disadvantage of feature subset selection or Feature Selection (FS) is that some features that may seem less important, and are thus discarded, may bear valuable information. It seems a bit of a waste to throw away such information that could possibly in some way contribute to improving model performance. This is where Feature Subset Ensemble (FSE) comes into play. It simply partitions the input features among the individual prediction models in the ensemble. Hence, no information is discarded. It utilizes all the available information in the training set, and at the same time not overloads a single prediction model with all the features, as this may lead to poor learning.

Ensemble feature selection is a fairly new approach. Most of the techniques use random assignment of features among the networks. There is still a space for improving the performance of the ensemble feature selection through intelligent assignment techniques or specific weighting techniques, and this is one of the issues investigated in the present research work.

1.4.4 Importance of Feature Selection Techniques

The problem of feature subset selection is concerned with finding a subset of the original features of a dataset containing only the selected features will generate a predictive model that has the highest possible accuracy. It is essential to select a subset of those features which are most relevant to the prediction problem and are not redundant (Hand & Till 2001),

(Hall 2000). Feature selection is a core to training dataset selection since one of the motivating factors for training dataset selection is to improve predictive accuracy through variance reduction.

1.5 GENERAL OVERVIEW OF CLASSIFICATION

Classification is one of the most frequently studied problems by DM and Machine Learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes).

With respect to supervised learning techniques, a prediction function is obtained by inputting observed independent variables into a classification program which "learns" from a training sample which is then used to predict the values of interest (outputs or dependent variable). In the classification literature, the outputs are categorical and defined as class labels, often with no meaning of their ordering (Johnson & Wichern 1988).

A major step after constructing classification rule is to assess its accuracy. The reason for that is a classifier cannot always perfectly distinguish new objects. In particular, in some cases, the features of whole classes are mathematically indistinguishable. In other words, high overlapping of the available measurements between regions of sub-population can result in poor prediction.

According to this, it is extremely convenient to measure the performance of a classification rule. To achieve this, comparison between actual classes with corresponding predicted classes is carried out in order to count misclassified objects.

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome.

Importance of Classification

- Classification is used for data analysis that can be used to extract models describing important data classes or to predict future data trends.
- Classification is a data mining used to predict group membership for data instances.
- Classification is the task of generalizing known structure to apply to new data while clustering is the task of discovering groups.

In recent years, in order to improve the overall performance of classification, ensemble classification approach has been used. So, the present research work uses ensemble classification approach.

1.5.1 Ensemble Classification

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to

over-fit the training data more than a single model would, but in practice, some ensemble techniques (especially bagging) tend to reduce problems related to over-fitting of the training data.

An ensemble classifier consists of a number of base classifiers and makes the prediction by combining the results of individual predictions.

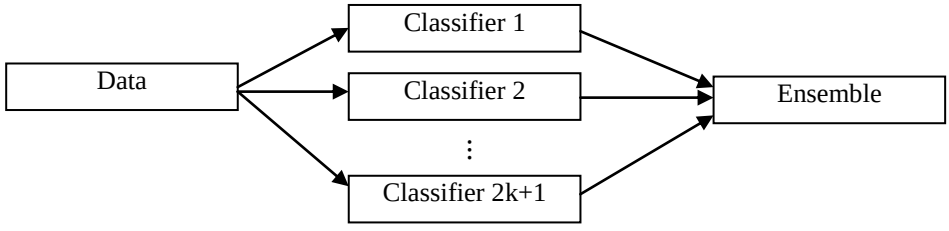

Figure 1.2 Ensemble classifier

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions. It has been applied to a wide range of real problems, such as object detection and recognition. As a multiple learner system, it tries to exploit the local different behavior of the base learners to enhance the accuracy and the reliability of the overall inductive learning system. There are also hopes that if some learner fails, the overall system can recover the error. Basically the effectiveness of ensemble methods relies on the independence of the error committed by the component base learner. From a general standpoint of view it is clear that the effectiveness of ensemble methods depends on the accuracy and the diversity of the base learners.

1.6 MOTIVATION OF ENSEMBLE CLASSIFIER

The main advantage of this ensemble classification is the reduction of variance and thus, the results are less dependent on peculiarities of a single training set. Moreover, it reduces the bias in which a combination of multiple classifiers may learn a more expressive concept class than a single classifier.

1.6.1 Homogeneous Ensembles

Homogeneous ensemble methods use the same base learner on different distributions of the training set, e.g. bagging and boosting. Heterogeneous ensemble methods incorporate different model types into the library of models, the idea being that different base model types can be both accurate and diverse.

1.6.2 Heterogeneous Ensembles

A heterogeneous ensemble is an ensemble with a set of base classifiers that consist of models created using different algorithms. The same combination functions that are used to create homogeneous ensembles can be used to create heterogeneous ensembles. The main difference lies in the methods used for creating the set of base classifier. The methods available for creating base classifier sets for homogeneous ensembles are modified so that models built from different classification algorithms can be combined to form a set of base classifiers. Currently, there is no clear choice on how to combine these base classifiers most effectively. Furthermore, there are open questions regarding which base classifiers are to be used and how they should be combined for optimal performance.

1.7 MOTIVATION OF RESEARCH

Exploring and analyzing large datasets have become an active research area in the field of data mining in the last two decades. There have been several approaches available in the literature to investigate the large datasets that comprise of millions of data.

Large databases have aggravated the rapid growth of data mining and knowledge discovery, that is, the search for the arrangement in large volumes of data. Whereas science and industry have leveled up their data

collecting activities, conventional data analysis investigation in statistics and machine learning has been comparatively slow to begin the challenge and much research is still based on relatively small data sets.

1.8 OBJECTIVE OF RESEARCH

The present research work aims to investigate the usage of various efficient feature selection and classification approaches for improving the overall performance. This work describes the application of various feature selection and classification techniques which helps to analyze the large datasets. The present research work also focuses on improving the feature selection and classification accuracy by selecting the best features in the large medical dataset. The work analyzes the medical data records for diagnosing various types of diseases. Ensembling is the main principle introduced in both feature selection and classification models for handling the high volume, complex medical records.

1.9 PROBLEM SPECIFICATION

The most important data mining approaches involved in this task are preprocessing, feature selection and classification. All the three approaches have their own importance in carrying out the task effectively. Most of the existing techniques suffer from drawbacks of high complexity and computationally costly on large data sets. In feature selection Insufficient removal of attributes may result in the huge losses of internal information among features. It degrades the performance of the classification.

1.9.1 Issues and Challenges to Handle Large Dataset

The main issues to handle large dataset are high dimensionality and size. Additional issues are multiplicity testing, feature selection, correlation structure, over-fitting; etc.,

High Dimensionality: The difficulties in modeling of high dimensional data are challenging. High dimensional data are sparse and make model fitting computationally intensive.

Size: The size of dataset will vary with the number of independent entities in the dataset.

1.10 RESEARCH CONTRIBUTION AND METHODOLOGY

The present research work mainly concentrates on classification of large datasets by selecting the features through hybrid and ensemble models. This research work focuses on hybrid and ensemble based feature selection approaches. Homogeneous and heterogeneous ensemble classification models are presented for improving the classification accuracy.

Hybrid feature selection processed data is carried out using Enhanced Genetic Algorithm integrated with Kernel PCA_SVM algorithm. In the hybrid feature selection approach, the option of three positive parameters such as σ, ε and C of KPCA SVM greatly influences the accuracy of classification in large datasets. Therefore, enhanced genetic algorithms are integrated with the proposed KPCA SVM model to optimize the parameter selection. A negative mean absolute percentage error (MAPE) is taken as the fitness function for evaluating the fitness value.

Then, in order to still improve the performance of the feature selection, ensemble feature selection approach is presented which includes Hybrid Genetic Particle Swarm Optimization (HGPSO) and Hybrid Artificial Bee Colony (HABC).

The present research work uses both homogeneous and heterogeneous ensemble classification approaches. In homogeneous

classification approach, Fuzzy k-Nearest Neighbor Algorithm (FKNN) classifier has been utilized. In heterogeneous classification approach, three classifiers namely Fuzzy k-Nearest Neighbor Algorithm (FKNN), Fuzzy Rule based Classifier (FRB) and (Adaptive Neuro Fuzzy Inference System) ANFIS classifiers have been integrated.

The present research work utilizes the ensemble model in medical diagnosis application. Moreover, in medical applications, accuracy of the classifier is an essential aspect. A high percentage of false negatives in the classification systems become a challenging issue that has to be sorted out through efficient classification models. Thus, the present research work focuses on developing an efficient classification model for the diagnosis of medical record.

1.11 ORGANIZATION OF THESIS

Chapter 1 provides an introduction to the data mining; data mining with preprocessing methods, feature selection methods, Importance of these methods and Problem specification, motivation and finally the research contribution is thoroughly explained.

Chapter 2, "Literature Survey" discusses about the existing feature selection methods, classification methods for large dataset analysis .The inference from the existing techniques and its limitations are also discussed in this chapter.

Chapter 3 deals with the detailed explanation of "Hybrid and Ensemble Model based Feature Selection Approach for Large Datasets". This work mainly focuses on effective feature selection results on large datasets. Two feature selection models namely hybrid and ensemble models are presented in this chapter. Hybrid feature selection model based on GA

integrated with KNN is presented. Ensemble feature selection approach discusses about the ensemble model with swarm intelligence approaches.

Chapter 4 discusses about, “Fuzzy based Ensemble Classification Model for large datasets”. This chapter presents homogeneous and heterogeneous classification models with fuzzy based approaches.

Chapter 5 discusses about the third proposed methodology namely, “FUZZY ENSEMBLE CLASSFICATION METHODOLOGY ON MEDICAL DIAGNOSIS”. This chapter investigates the proposed methodology for medical diagnosis with different bioinformatics datasets.

Chapter 6 presents a performance evaluation of the proposed approaches with different bioinformatics datasets. The performances of the approaches are validated under various metrics.

Chapter 7 concludes the thesis with the findings of feature selection with hybrid methods and classification for larger dataset. This chapter also discusses about the scope for future improvement.