CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION TO MACHINE LEARNING

Knowledge Discovery from Data (KDD) is data mining [1] and is used in various fields such as information mining (IT), finance and medicine. Rapid development of pre-existing databases helps to turn data into useful facts and knowledge. As huge data is gathered at enormous speed with multi-dimensional and distributed way, conventional models require efficient machine learning model for data analysis. However, these models take high computational time for pattern discovery in order to extract hidden information in the large databases. Data mining or knowledge discovery from databases (KDD) means extracting very thought-provoking and significant patterns and relationships between the information stored on large databases. These results are recycled for different objectives and for different areas, ranging from organizational management, product marketing, research and development, disease diagnosis, data security and MIS decision-making. Such choices are defined by enormous amounts of data that are related in various ways to each other. Because of the massive data and high data growth rate in the medical field, human analysis skills are inadequate. Because biotechnology is developing at a high rate of growth, more and more biological data is being collected and made available for analysis. The importance of developing new techniques to extract knowledge from it also increases when bio-molecular data grows significantly. The healthcare environment is generally seen as rich in information, but unfortunate in the distribution of knowledge [2]. Within the various healthcare systems, the rich data is readily accessible. However, there are very few analytical tools available that are efficient and able to extract relationships that are hidden between data and data trends. Mining of data and exploration of information help many applications to find information in the areas of medical science and other fields of science. The knowledge that has a high value is identified from the applications. These applications are techniques for the mining of medical hospital data. Through this research, we intend to observe the possible use of data mining techniques that are the basis of classification and the decision tree of massive volumes of health data.

1.2 IMPORTANCE OF CLASSIFICATION

Classification is an attempt to predict the class label for each case in the data accurately. This is how a classification task begin with known class assignments. So, a credit risk prediction model for many loan applicants over the period of time can be developed. Besides or alternatively, it might capture education, employment history, home ownership or rental history, years of residence, number of investments, and so on. Credit rating would be the target, other attributes would be the predictors, and customer data will become the cases. Classification model can be used to find relationships between the values of predictors and the values of the target variable. Some classification systems use different techniques when discovering relationships between objects. A model of these relationships can be applied to other data set in which the classes cannot be allocated beforehand. Classification models are checked by contrast by comparing their predictions of target values with known target values in sample data. The model is typically divided into two parts; one for building the model and the other to test the model. Classifier models produce classifications and assigns probabilities for each case. In the same way, that type of models predicts the probability of each classification for each customer.

1.3 MOTIVATION

Identifying interesting or hidden patterns play a vital role in many real time applications, such as decision making, business intelligence, and data mining. For example, an anomalous transaction of a credit card may imply unauthorized usage. Data mining anomaly detection techniques have been used in the past to detect fraud usage of credit cards, medical disease prediction and network anomaly detection. The data base sizes were about 500 to 1000 items. The accuracies obtained were 50 to 75 %. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis.

With the emergence of large data collected from real-world scenarios it is important to apply the anomaly detection techniques for the large data also. The conventional anomaly detection models are inefficient and infeasible, as the size and number of instances are large. As the size of network data increases, the risk of cyber-attacks on the complex networks are also increases accordingly also they not give better accuracy. In order to improve accuracy new methods are required, In this thesis work

several modified machine learning approaches have been used to attain higher level of accuracy in classification.

The extraction of hidden information from heterogeneous database is one of the major issue in real-time applications. In data mining research, the data mining algorithm for both high-dimensionality and cardinal data was considered one of the challenges. The large number of data generated with different heterogeneous features are used in the conventional models for pattern discovery

Selection of subsets of features or sub-set attributes is useful in different data mining and machine learning algorithms. The advantage of selecting sub-set functions is multiple folds. These advantages motivated us to address this issue. The challenge is to find the key features in order to eliminate from the real-world data to optimize classification performance. This leads to high training accuracy and low testing accuracy; the over-adjustment problem is known. This problem occurs because of the fact that the data have irrelevant and noisy features. These algorithms also do not allow for useful decisions. Therefore, the choice of features is needed to decrease over the use of large data. Feature selection methods can solve this problem by feature subset using ranking methods.

1.4 PROBLEM STATEMENT

1.4.1 Research Gaps:

The Following Gaps Are Identified

- Traditional classification models failed to find the anomaly patterns prior to pattern analysis. The primary challenge in constructing the patterns the traditional classification models between the normal class and anomaly class.
- Most of the conventional techniques are based on binary classification for tree construction anonymity. As the levels of decision tree increases traditional model require high computational memory with high error rate.
- Traditional models use homogeneous data type for data classification with limited data features. Mixed attributes, High true negative rate pose lot problems. They result in High classification error rate due to imbalance property. Traditional models perform decision making on the positive associated pattern.

Most of the conventional techniques are based on data size with limited number of attribute sets during training and testing phases. Traditional models degrade performance on high dimensional features for pre-anomaly detection. Difficult to check

the attribute based anomaly detection on large datasets such as somatic cancer and network datasets

The main focus of this research is to optimize the prior probabilities of medical data feature classes and gene-disease prediction for improving the true positive rate of the scalable classification algorithm.

1.4.2 Solution

Therefore, it is necessary to use modified data mining techniques to overcome the problems. This thesis work proposes the following modified techniques to improve accuracy in anomaly detection thereby detecting diabetics, cancer deceases, correct type of glass used in criminological investigation and better attack detection.

1. **Probabilistic Feature Engineering**
2. **Classification Accuracy Enhancement through Meta-learners**
3. **Re-engineering of ensembles for real world datasets.**

The development of above models is described in subsequent sections. The proposed work is summarized in the form of following objectives.

1.5 RESEARCH OBJECTIVES

The objective of the research is to find the anomaly in the numerical attributes by classifying the data using the advanced decision tree models. The main objectives of the research are

1. Improving Classification accuracy using **statistical control chart based extreme outlier removal model combined with modified decision tree model** for distributed data with modified entropy and information gain to better detection of diabetes, correct type of glass, network attacks, cancer decease and credit risks in the credit dataset.
2. Implementing **Relevant feature selection model with multivariate analysis** for high dimensional dataset for more efficient classification leading to better detection of diabetes, network attacks, cancer decease and credit fraud datasets.
3. **Improving Classification accuracy with Ensemble based Partition classifier model** with **large number of feature sets** leading to better detection of diabetes, network attacks, cancer decease and credit risks in the credit fraud datasets.

1.6 DATA DESCRIPTION

Many studies have contrasted neural networks with other methods of scoring. In a study that compared discriminant analysis, logistic regression and neural network models, the neural network model had an overall efficiency rating that was higher than logistic regression and support vector machine. Credit scoring is still very limitedly known by the world. Although credit scoring started in 1941, until the turn of the century, the use of credit scoring was limited to client evaluation. The uses of credit scoring have expanded to different areas in the short period of life. The following section deals with the review of studies in various fields concerning the application of credit scoring. The first credit scoring application, who recognized the need for a numerical value that could quantify the applicant's creditworthiness and assist the lender in the process of subscription. The tool's prominence has also increased with increased pressure on the banking sector to generate more credit facilities. Wider product ranges have forced creditors to customize their models. The use of credit scoring has historically been restricted to the assessment of home loans, mortgages, and consumer credit. Lending to small businesses, however has been viewed by most lenders as a risky game. Banks preferred to lend with large valued collateral, accounted financials, and proven credit history to large enterprises that processed systematic future plans. Many small businesses find it very difficult to get loans from lenders. A major hindrance has been experienced by lenders because they have had no place to grow. Lenders are looking for opportunities for their loan customers' repayment history to improve. Banks and other formal lenders are now targeting small businesses whose financial needs have been met primarily by local money lenders. Therefore, one of the lender's important priorities is to look out for creditworthy borrowers. Credit scoring has proven to be an effective tool for assessing the associated credit risk and reducing lending costs to small business borrowers.

Cancer research is one of the major fields of medical research. Understanding the likelihood of cancer developing is an important feature to improve the treatment of cancer. Microarrays can be used to diagnose different types of cancers and to predict it. For example, research into the classification of lymphoma, leukaemia, breast cancer, and liver cancer has employed gene expression data very effectively. Microarray technology provides a new tool for automating the diagnostic work and improving the precise traditional diagnosis techniques. The expression of thousands of genes can be

examined at once with microarrays. Higher expression testing of certain genes can help cancer predict. The problem in the analysis of microarrays however is that gene expression data are ultra-highly dimensional (microarray image). The high dimension of microarrays makes it extremely difficult to process them and their complexity in time and space. Therefore, it is important to reduce the data dimensionality before further processing in order to make processing microarray feasible. The t statistics are directed to the medium differences between the interclasses and inverse to the standard deviations of the interclasses. Small standard deviations in the intrinsic classes and a large interclass mean difference show a good class gene (small p-value). Based on the overlap of distributions, a p-value is determined. The microarray gene classification of cancer is a major problem for classification problem.

KDD99dataset: MIT Lincoln Laboratory conducted a reckon for evaluating intrusion detection tools in 1998-99. This research work was supported by DARPA, an intrusion detection community for further attacks evaluation. This KDDCup'99 data is most commonly used public intrusion detection dataset. The KDD99 data consists of both labeled and unlabeled instances.

1.7 ORGANIZATION OF THESIS

Chapter 1 Describes the introduction to anomaly detection models, different classification approaches to numerical and mixed type of attributes. In this chapter, motivation, research gaps, problem statement and research objectives are discussed.

Chapter 2 Presents the survey on different types of anomaly detection models, classification models, and statistical analysis methods on medical, credit card databases. In this chapter, the problem of different ensemble learning models are also discussed on different applications.

Chapter 3 Provides information about Design Methodology of statistical anomaly detection model on the numerical type of attributes for classification problem. In this chapter, a hybrid anomaly detection model is proposed for decision tree classification problem on different databases such as credit card, kdd and medical datasets.

Chapter 4 Presents the feature selection based classification model on high dimensional mixed feature space. In this chapter, a hybrid feature selection model is proposed for optimized decision tree classification problem on different datasets.

Chapter 5 Present the hybrid feature selection based ensemble learning model on large databases and high dimensional feature space. In this chapter, an advanced ensemble learning model is proposed in order to improve the overall prediction rate on different types of application datasets.

Chapter 6 Presents the conclusions and future scope of the research work.

1.8 SUMMARY OF THE PROPOSED FRAMEWORK

In this work, a hybrid framework is designed and implemented on the different real-world datasets such as medical dataset, network dataset and credit card dataset. Proposed framework is implemented in three phases. The framework will have 3 phases 1. Probabilistic Feature Engineering 2. Classification Accuracy Enhancement through Meta-learners 3. Re-engineering of ensembles for real world datasets. In the first phase, probabilistic feature engineering module is developed in order to find the essential key features on different datasets. First phase is implemented and discussed in the chapter 3. In the second phase, a meta-heuristic based classification model is implemented on different datasets in order to optimize the error rate and accuracy. Second phase is implemented and discussed in chapter 4. Finally, in the third model, a hybrid ensemble learning model is developed on the different real-world datasets. In this phase, real-world datasets such as networking data, medical data and credit card case study data are used to evaluate the experimental results. This phase is implemented and discussed in chapter 5.