CHAPTER 1

INTRODUCTION

Information technology is developing day to day at a tremendous pace. This tremendous development leaves behind a large volume of digitized data. If this massive data is analyzed properly, then useful knowledge can be gained to make the smart decision. Data Mining is a technique that can execute innovation activity in a coherent manner and find out the rules and modes to obtain the necessary knowledge that help in making decisions. The arena of data mining is a unique area that plays a significant role in various fields such as Education, Medicine, Business, Finance, Data Analysis, Marketing, and Stock exchange. Data mining is also called as Knowledge Discovery in Databases (KDD), data dredging, knowledge extraction, knowledge mining from data, data archaeology and pattern analysis (Millie Pant et al. 2015). Knowledge Discovery Databases (KDD) is the method of extracting or mining, information from huge quantities of data (Tipawan and Tuamsuk 2012). As Machine Learning (ML) algorithm uses statistical or pattern methods to discover models, patterns, and other regularities in data it is widely applied in data analysis to extract the knowledge from data and provides tools for analyzing the large size data.

1.1 DATA MINING

Data mining uses several data mining tools for analyzing data from various dimensions, to classify, and to review the data. It is the procedure to
discover patterns and relationships hidden in the data. Figure 1.1 shows the procedure of information discovering in data mining as collective action. The result of data mining helps in decision making, understanding the behaviors of customers, predicting the behavior of clients, anticipating the need of consumers, reducing the cost and improving sales.

![Figure 1.1. KDD Collective process](image)

Data mining is a method identified as knowledge discovery from databases. Data mining brings different types of people working as computer researchers, arithmeticians, statisticians, bioinformatics researchers and data analyzers. Data mining is used in areas like artificial intelligence, machine learning, pattern recognition and information retrieval. Data mining utilizes precisely straightforward statistical procedures, or practice extremely refined data analysis (Arun K Pujari 2005).
Figure 1.2 shows the extensive model of the data mining task. Data mining is applied in several applications that deal with humungous of data. In the area of market analysis, data mining helps in pricing products and in finding out the behavior of buyers. In Customer Relationship Management (CRM) the transaction data is mined to know the customer preference so that new customers could be attracted and the existing customers are retained. In the banking sector KDD technique is used in segmenting the profitable customers and in retaining them. Data mining is also helpful in disease diagnosis, disease treatment optimizing, gene finding and protein finding.
1.2 ARCHITECTURE OF DATA MINING SCHEME

The components of a typical data mining architecture are shown in the figure 1.3.

- The database, data warehouse, and repositories: Data cleaning, selecting, filtering and data integration methods can be achieved in this phase.
- Database or data warehouse server: Database server is liable for fetching the information based on the request from the data mining engine.
- Data mining engine: Data mining engine performs tasks association analysis, characterization, classification, clustering, evolution, and deviation analysis.
- Pattern evaluation: This phase communicates with the data mining engine to search for attractive patterns based on the user queries.
- User interface: Interact with users and the data mining engine, permitting the user to connect through the section by identifying a data mining query or task. This module permits the user to scan data warehouse or data structures and database, assess extracted patterns, and visualize the patterns in various actions.

### 1.3 DATA PREPROCESSING

The methods used to gather data from different sources are not properly controlled. The data collected using such methods might contain useful as well as unwanted or garbage data. Hence the need for converting the data into a useful data is unavoidable. Data preprocessing is one of the significant phases in data mining techniques which analyze and convert data into usable information. The forms of data preprocessing is shown in the figure 1.4. The data taken for analysis is subjected to three different types of studies. Initially the study of incompleteness, which address the problems like missing feature values and missing features of interest. Secondly, the noisy study where the errors or outliers analysis are taken care and thirdly the study on inconsistency. At the data preprocessing the quality of data is improved, thereby helping to improve the accuracy and efficiency of the subsequent mining process.

#### 1.3.1 Data cleaning

The dataset might contain noisy data, outliers or inconsistent data. This dirty data can cause some confusion in the data mining process or yield an
unproductive result. Therefore the data cleaning process becomes an unavoidable process in building a dataset free from noise, redundancy and inconsistency. Data cleaning process fills the missing values, smooth out the noise, identifying the outlines and removing the inconsistency. The process of filling missing values uses several techniques such as ignoring the tuples, physically filling the missing values, using an overall persistent to fill in the

**Figure 1.4. Forms of Data Preprocessing**
missing values, and use the feature means. Likewise smoothing noisy data is an additional data cleaning method. Noise is an accidental error or alteration in a measured variable. Some of the smoothing methods are binning, regression and clustering (Jiawei Han and Micheline Kamber 2007). The smoothing method is used to recognize or eliminate outliers and to determine inconsistencies in the dataset.

1.3.2 Data integration

The massive data for knowledge mining is not constructed from a single data source, but extracted from various sources of information. The source data are constructed by extracting data from different databases or data cubes, by integrating multiple databases and files. The information accumulated from these sources, forms an intelligent data store and is presented with the data mining engine or machine learning tools (Miriam 2005) for extracting knowledge. As the data store is constructed from extracts from different sources of data it is found to contain redundant data. The data redundancy is detected using correlation method chi-square and is eliminated.

1.3.3 Data transformation

In data transformation, the information is converted or modified into models suitable for mining. The various data transformation methods are smoothing, generalization, attribute construction, normalization, and aggregation. There are three main approaches for data normalization. They are categorized as z-score normalization, normalization by decimal scoring, and minimum-maximum normalization (Jiawei Han and Micheline Kamber 2007).
1.3.4 Data reduction

Data reduction method possibly enforces to achieve a shortened representation of the source dataset, particularly far lesser in dimensions, however sustaining the reliability of the original information. However, the compact dataset is highly efficient to achieve the similar analytical results. Data reduction uses strategies such as feature subset selection, data aggregation, dimensional reduction, numerosity reduction, discretization and notion hierarchy generation. The data reduction strategies are data cube aggregation, dimension reduction, data compression, numerosity reduction and discretization and concept hierarchy generation. The numerosity reduction method can be nonparametric or parametric. The parametric technique is applied to calculate the information that contains regression and log-linear models. Nonparametric technique includes methods like histograms, clustering, data cube aggregation and sampling techniques (Jiawei Han and Micheline Kamber 2007).

1.4 DATA MINING TECHNIQUES

There are several data mining methods that have been developed and used in data mining. The following section examines few significant methods of data mining techniques:

1.4.1 Association

The association rule mining technique discovers all robust association rules from the frequent itemsets that exists in the dataset. The research in the field of association rule mining deliberates on how to discover frequent itemsets
efficiently (Li Shen et al. 1999). The frequency and the number times of occurrence of itemsets are the factors used in determining association rule. Association rule mining technique generates a very large number of rules which leads to processing overhead and overfitting. These overheads are addressed by pruning technique. Pruning technique discards unnecessary and redundant rules and the size of the resulting rules set is far smaller than the rules resulting out of normal association rule mining (Jiuyong Li et al. 2002). An association finds applications in market basket analysis, item placement, fraud detection, medical research, process re-engineering.

1.4.2 Classification

Classification is a data mining technique distributes items in a collection to known fixed target categories or classes. The simplest form of classification technique is a binary classification where the target categories are two. Classification divides the historical data into two data sets: one for building the model; the other for testing the model. Classification process involves two steps. First step, also known as learning phase learns the relationship between the attribute sets and class label of the given training dataset and builds a model. Later this model is tested for accuracy using test dataset. Finally the model is used in the classification of a new or real-time dataset.

A standard classification algorithm includes naive bayesian classification, decision tree learning, k-nearest neighbor, neural network and support vector machine. Though classification designs a classifier model the time spend on the learning process is an overhead which cannot be avoided. Data Mining can be used for knowledge discovery of interest in Human
Resources Management (HRM), remote sensing images, disaster weather forecasting, medical diagnosis and so on. Likewise, the definite knowledge growth is suitable for several types of research as because the decision-making method largely depends upon the success of the classification technique being developed (Soumadip Ghosh et al. 2014).

1.4.2.1 Data classification working model

The data classification working model is processed in a two-stage such as learning step and classification step as shown in figure 1.5. In the learning step, the training information is examined by a classification procedure, class label feature and learned model classifier is denoted in the method of

![Figure 1.5. Data classification working model](image-url)
classification rules. Training tuples are the individual tuples that make up the training set.

A tuple $V$ is denoted by a n-D feature vector $V = (V_1, V_2, \ldots, V_n)$ and Attribute denoted $A = (A_1, A_2, \ldots, A_n)$. The second stage in classification evaluates the precision of the classifier model using the test data. If the precision result is found satisfactorily, the classification rules can be used for the classification of fresh information tuple (Jiawei Han and Micheline Kamber 2007).

1.4.2.2 Decision tree

Decision Tree is similar to a flowchart structure. The decision tree is a widely used classification technique because it can handle high dimensional data, makes the process of learning and classification simpler, can easily build a predictive model, and yield a very accurate result. The node at the top of the tree is called the root node. The decision tree starts branching out from the root. Every internal, or non-leaf node in the decision tree represents some test on an attribute and the outcome of the test decides the branching. The quality attributes such as Information Gain, Gain Ratio, Gini Index etc. helps in the node selection from the top level.

1.4.2.3 Decision Tree Induction Algorithm

In the last three decades, researchers have developed many algorithms that have enhanced the accuracy, performance and the ability to handle various types of data. Ross Quinlan (1980) introduced a decision tree algorithm called
as ID3 (Iterative Dichotomiser). Future Ross developed C4.5, an alternative to ID3. Both C4.5 and ID3 implement greedy method. The decision tree algorithm assembles trees in a top-down recursive divide-and-conquer technique, and there is indeed no backtracking. In the decision tree method node with highest information gain (entropy reduction) is generally chosen for branching. Figure 1.6 is the decision tree model used in banking sectors in approving for different types of loans.

![Decision Tree Model for Loan Dataset](image)

**Figure 1.6. Decision tree for loan dataset**

### 1.4.2.3.1 Tree Pruning

The decision trees generated by the ID3, C4.5 are accurate and efficient, but many of the branches in the large tree reflect anomalies in the training data
due to noise. To identify and remove the branch that does not improve accuracy pruning technique is used. The classification accuracy and the computational complexity of a pruned tree is found to far better than the tree generated by ID3 and C4.5. Moreover, after pruning the tree size gets reduced. Pre-pruning and post-pruning are the two types of decision tree pruning techniques. Tree pruning might increase the performance of classifiers by removing error-prone element (Eibe Frank 2000).

1.4.2.3.2 Tree Pruning Methods

**Post-pruning:** Post-pruning decision tree technique yields a small and an accurate tree in a shortest span of time. The post-pruning efficiency depends on the significance test. In post-pruning process two types of significance tests are followed. They are parametric test and non-parametric test. Non-parametric tests have the advantage that they do not make assumptions about the specific functional form of the test statistic’s distribution. If the appropriate significance level is chosen, then a tree resulting out of seems to be insignificant. However, as the permutation tests try all possible permutations with the given data it is found to be more computationally expensive when compared to parametric test.

**Pre-pruning:** Pre-pruning as compared to post-pruning does not prune existing branch instead the subset of the existing branch is tested for any increase in accuracy. If a sub-tree shows improvement in accuracy, then the branch is chosen for expansion. Otherwise the expansion of the existing is suppressed. When post-pruning and pre-pruning are compared post-pruning is found to be a more efficient pruning method.
1.4.2.3.3 Extracting classification rules from decision trees

A rule-based classifier is a technique for classifying records using a collection of “if ... then ...” rules. Figure 1.7 shows the decision tree which uses loan datasets to predict risk or safety in sanctioning loan.

```
| children <= 1 |
| children <= 0 |
| marital = NO |
| loan = NO: YES (45.0/3.0) |
| loan = YES |
| deposit_act = NO: YES (11.0) |
| deposit_act = YES: NO (21.0) |
| marital = YES |
| deposit_act = NO |
| loan = NO |
| salary <= 32617.2 |
| age <= 39: NO (10.0/1.0) |
| age > 39: YES (5.0/1.0) |
| salary > 32617.2: NO (22.0) |
| loan = YES: YES (26.0/3.0) |
| deposit_act = YES: NO (120.0/12.0) |
| children > 0 |
| salary <= 15538.8 |
| age <= 39: NO (24.0/2.0) |
| age > 39: YES (2.0) |
| salary > 26649.8: YES (122.0/5.0) |
| children > 1 |
| salary <= 41515.3: NO (132.0/12.0) |
| salary > 41515.3 |
| children <= 2: YES (53.0/5.0) |
| children > 2 |
| salary <= 53377.3: NO (20.0/2.0) |
| salary > 53377.3: YES (9.0) |
```

Figure 1.7. IF-THEN rules for classification
The data signified in decision trees might be extracted and characterized in the usage of IF-THEN rules for classification. Every path from the root to a leaf node represents a rule. An item is assigned to the class that meets the conditions along the path from the root to the leaf it reaches.

1.4.3 Cluster Analysis

The unsupervised machine learning technique, ‘clustering’ groups the related objects. Cluster analysis employs statistical methods to dig out clusters from the underlying data. The clustering is also a type of classification, but it differs from it in the process of labeling the objects. Clustering creates labels for object class labels which are derived from the data. In classification, unlabeled objects are assigned the names of the model created for that class of data. Clustering could be broadly classified as hard clustering and smooth clustering.

The various types of clustering algorithms are classified as partitioning technique, hierarchical technique, constraint-based technique, density-based technique, model-based technique and grid-based technique. Clustering analysis is applied in various applications, like market research, biological field, pattern recognition, data analysis, categorizing documents on the web for information detection, image processing, credit card fraud detection, classify of regions of related land use of a ground database, and so on.
1.5 MOTIVATION FOR STUDY

Earlier studies of feature selection have addressed the feature selection issue in high dimensional data by applying hybrid methods with two feature selection algorithms (Hui-Huang Hsu et al. 2011, Afef and Mohamed 2016). From the literature survey, it is observed that most of the feature selection methods focus on classification accuracies and did not show most interest in addressing the time complexity (Haitao Liu et al. 2013). Moreover, it is also found that only a specific type of datasets has been used, such as UCI machine learning datasets or medical datasets or banking datasets and so on (Wenbin, Wenhao 2015) and multiple datasets are avoided. This motivates why not to proceed with three feature selection techniques to select the best feature subsets and address the computational time with multiple datasets. The AdaBoost with the decision tree classifiers are fairly used in many boosting algorithms; consequently, the decision trees is non-linear, which inspired and motivated to go with other classifiers like Random forest, C4.5, and Naïve Bayes classifiers.

1.6 PROBLEM STATEMENT

High-dimensional datasets contain a broad feature set which is desperate to learn and produce a subset of features. It also worsens the overall performance of the learning algorithms. Issues related to feature selections are irrelevant and redundant features in the original dataset, decreases the learning speed, efficiency, and accurateness of the learning algorithm. The training time gets extended, and the probability of error is also higher when the features are outside the limit. Datasets are having a large number of features are frequently occupied with hundreds or thousands of features which provide a possible discrimination for classification tasks. On the other view, they may weaken
classification performance by reason of the inadequate number of training data. The leading factors that influence the boosting algorithm are the ‘strength’ of the ‘weak’ learners, the noise level in the data, and the dimensionality of the data. The feature selection mechanism is critical to discover a subset of the feature because the entire existing features may not be used for ranking, but only a lesser percentage of these features may need for the classification task.

### 1.7 OBJECTIVE OF THE RESEARCH

The proposed research work aims to build the hybrid feature subset selection for multiple datasets using various boosting algorithms for decision tree based classifiers.

**Specific objective**

- To design hybrid model by combining the subsets to form an aggregate subset.
- To develop an effective Adaboost algorithm to reduce the error rate by weighting the features.
- To develop a hybrid feature selection model and ensemble model to analyze the classifier performance.

### 1.8 ORGANIZATION OF THE THESIS

Chapter 2 provides a general review of the literature on hybrid feature selection approaches, boosting algorithms.
Chapter 3 describes the methodology in each phase for feature selection methods.

Chapter 4 explains the implementation of hybrid feature subset selection for decision tree based classifiers.

Chapter 5 explains implementation of ensemble methods of effective AdaBoost algorithm with decision tree classifiers.

Chapter 6 describes the merging method of the hybrid feature subset selection to combine with ensemble methods in meta-classifiers.

Chapter 7 concludes the work by explaining the importance of this research, a summary of the findings, and intends some suggestions for future research in this area.