AN APPROACH FOR IMPROVING ACCURACY OF PREDICTION USING ENSEMBLE MODELING

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Mansi Gera

(Roll No. 801332015)

Under the supervision of:

Dr. Shivani Goel

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"*An Approach to Improve the Accuracy of Prediction using Ensemble Modeling*"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of ***Dr. Shivani Goel*** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature: Mansi Gera

(Mansi Gera)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Shivani Goel

(Dr. Shivani Goel)

Assistant Professor

Computer Science and Engineering Department

Countersigned by

(Dr. Deepak Garg)

Head

Computer Science and Engineering Department

Thapar University

Patiala

(Dr. S. S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude to **Dr. Shivani Goel, Assistant Professor,** Computer Science and Engineering Department, Thapar University, Patiala for introducing me to data mining and for all her guidance and support. This thesis work was enabled and sustained by her vision and ideas. I have been amazingly fortunate to have an advisor like her who gave me the freedom to explore new ideas on my own and at the same time she guided me to recover when my steps faltered. Her patience and support always helped me overcome many crisis situations and successfully complete this dissertation.

I am also thankful to the HOD **Dr. Deepak Garg** sir and entire faculty and staff of Computer Science and Engineering Department (CSED) and my friends who devoted their valuable time and constantly supported me in all possible ways towards completion of this work. I thank all those who have contributed directly or indirectly to this work.

Lastly, I would also like to thank **my parents** for their years of unyielding love and encouragement. They have been my backbone and have supported me in all walks of life.

Mansi Gera

Mansi Gera

(801332015)

ABSTRACT

In general terms classification can be divided into two steps. First one is learning step which consists of the predetermined set of classes or concepts. Second step involves testing, in which data sets are being tested for the verification. And after the system is trained with data it can be used for further analysis to be done in future so that future events can be predicted in advance. For different applications we need to apply these models to predict and note the accuracy given by each model. The main aim of the research here is to make such a system which has more accuracy as compared to what previous systems are giving. So to implement this type of system, a hybrid approach is used i.e. ensemble of classifiers. It is necessary that one does not weigh one model purely. Other models or methods can also give more efficient results. In doing so, give weightage to each method and combine these methods to reach final destination that is most informed one. There are a large number of models available which are used for classifying the data into various class labels. That is also known under various other names, such as multiple classifier systems, committee of classifiers, or mixture of experts. The basic aim of ensemble based systems is shown to produce favourable results compared to those of single-expert systems for a broad range of applications and under a variety of scenarios. There are various procedures available through which the individual classifiers can be combined. These procedures are called as combination rules. Each rule has its different functionality which will work according to requirement and application where it is applied. So the study is carried on the prediction by applying methods and ensemble that with variable seed values. The experiment is carried out on the k-fold validation to check the consistency of the system.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Knowledge Discovery in Databases (KDD)

Knowledge discovery is the most popular research area in the field of scientific and industrial community and basically used for finding patterns which gives knowledge. Knowledge discovery process consists of various steps that are business understanding, data understanding, data preparation, modelling, evaluation and deployment. The most essential step in this whole process is modelling or data mining. The techniques of data mining will help in making some intelligent decisions. Process of extracting information to make it useful, novel, understandable is basically the data mining or knowledge discovery in databases. Conventional database systems are often called as OLTP (Online Transaction Processing) systems which are designed for day to day running applications to obtain maximum throughput.

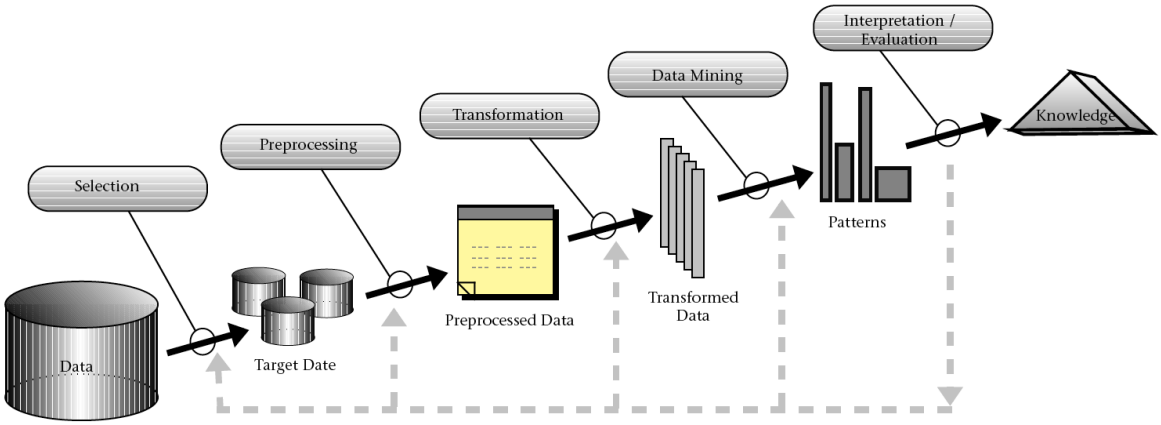


Figure 1.1 KDD Knowledge Discovery in Databases Process [1]

On the other hand, data warehouse is just the collection of historical data which may or may not contain the whole information. As in some applications data mining analysis requires complete information of the customers but that may not be saved in the database of data warehouse. So it varies from application to application. Basically data warehouse is used for storing the summary of the information and it consists of

the OLTP systems so that it can support the queries of customer, reports and analyzed information. The basic reason for using data mining is due to several reasons:

- There is huge growth in the OLTP data. As day to day data rate is increasing and large storage of data is required.
- Data from cards like online shopping sites and data from mobile phones are increasing day by day. There are lots of transactions being processed through debit cards and credit cards. Earlier less number of customers are using online transactions but now a day's everyone is using this.
- Development in the data available by websites which has become the biggest source of data.
- Development in the area of banking transactions, immigration transaction, utilities transactions. Based on these improvement in these areas data mining concept becomes popular.

By applying various techniques the new patterns will be obtained which gives us new knowledge. The techniques can be predictive or descriptive this all depend upon the data mining task and application which is to be solved. To make predictions about a certain characteristics of new data the predictive data mining techniques is being applied to dataset. If the technique which is used to make descriptive relationships not only is related to one specific characteristic of data but for all, then descriptive model is being build.

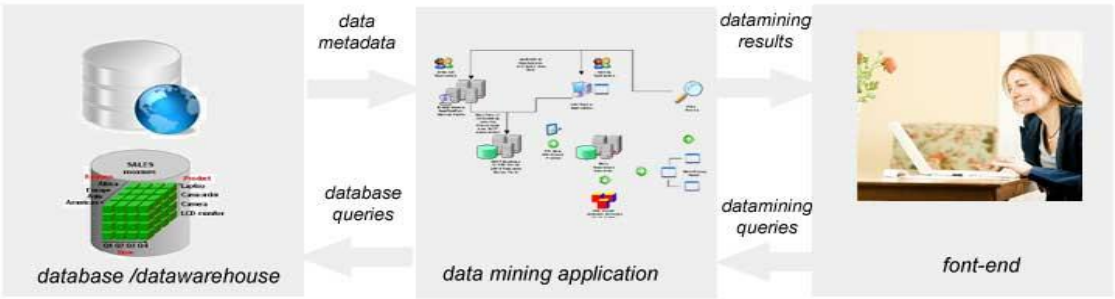


Figure 1.2 Architecture of data mining

1.2 Techniques of data mining

To analyze large amount of data, data mining came into picture and is also called as KDD process. To complete this process various techniques developed so far are explained in this section. The work presented here is in data mining in which useful outcome is predicted from large database. In most of the cases, it uses already built tools to get out the useful hidden patterns, trends, sequence and prediction of future can be obtained using the data mining techniques. Data mining involves model to discover patterns which consists of various components. Any data mining algorithm consists of following components:

- **The model**: The model is defined by the functions, e.g. classification, clustering, association, regression etc. To represent them as neural networks, the parameters are to be determined from the collected data.
- **The preference criterion**: The preference is all upon the data and parameters taken for data. Too many degree of freedom are to be constrained of data space and smoothening is required to be done to avoid over fitting of data.
- **The search algorithm**: It is used according to specification of algorithms for finding parameters and particular model given with data, precision and models.

Broadly the data mining is divided into different techniques classified as follows:

1.2.1. Classification:

It is one of the data mining techniques which are useful for the prediction of group membership for data instances. In general terms classification can be divided into two steps. First one is learning step which consists of the predetermined set of classes or concepts. These concepts are built by analyzing the previous records and training database instances. Second step involves the testing in which data sets are being tested for the verification. The accuracy of the model designed is being checked. In the end, this model acts as decision making classifier. There are various techniques to be followed for classification which are: decision tree, Bayesian methods, rules based algorithms, and neural networks etc.

- Decision tree induction: From the class labeled ‘tuples’ the decision tree is build. Decision tree is tree like structure in which there are internal node, branch and leaf node. Internal node specifies the test on attribute, branch represents the

outcome of the test and leaf node represents the class label. Decision tree handle high dimensional data.

- Rule – based classification: It is represented by set of IF- THEN rules. First of all how many of these rules are examined and next thing is about how these rules are build and can be generated from decision tree or it may be generated from training data using sequential covering algorithm. Expression for rule is:

IF condition THEN conclusion

- Classification by backpropagation: Backpropagation is a neural network learning algorithm. In neural network there is set of inputs, weights and output is associated to them on the basis of which the result can be obtained. Neural network learning is often called connectionist learning as it builds connections. It is feasible for that application where long times, training is required.

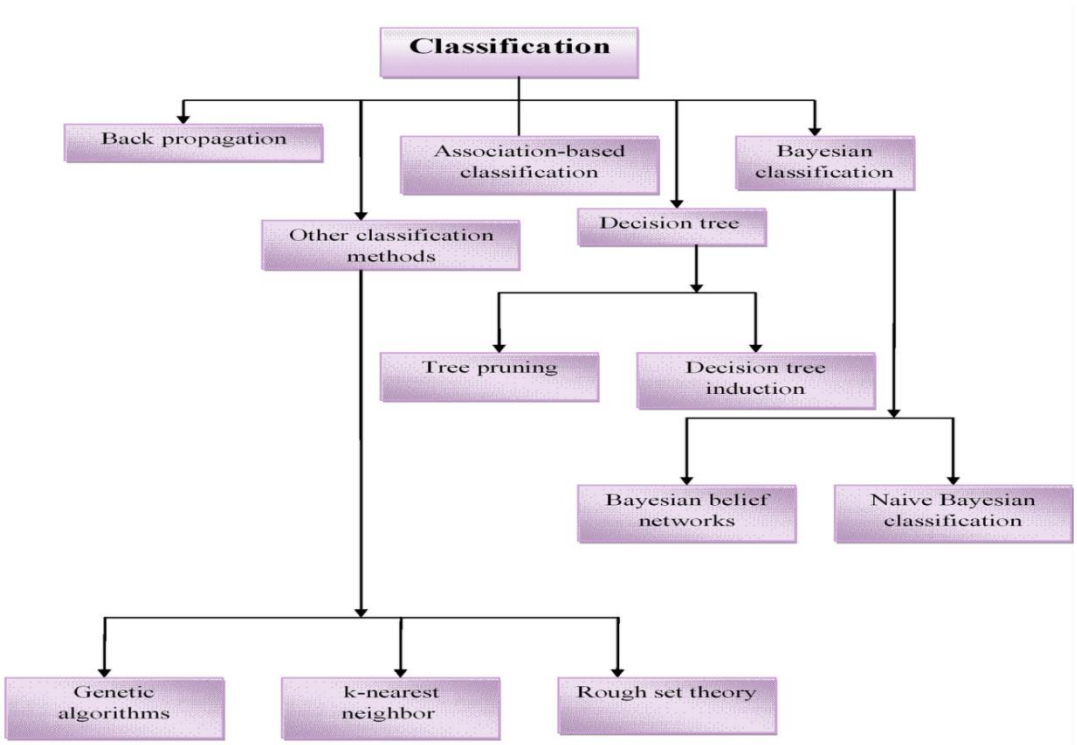


Figure1.3. Methods of Classification

- Lazy learners: Eager learner is the form in which generalization model is being developed earlier before new tuple is being received for classifying. So all these decision tree induction, classification by backpropagation and rule based classification are examples of eager learner. In lazy learner approach when given a training tuple it simply stores it and waits until a test tuple is given.

1.2.2 Clustering:

Unsupervised classification is called as clustering or it is also known as exploratory data analysis in which there is no provision of labeled data. The main aim of clustering technique is to separate the unlabeled data set into finite and discrete set of natural and hidden data structures.
Broadly clustering has two areas based on which it can be categorized as follows:

- Hard clustering: In hard clustering same object can belong to single cluster.
- Soft clustering: In this clustering same object can belong to different clusters.

Given there is set of input patterns Y= {y1,……,yi,……..yN},where y_i= (yi1,…..yid)T € R^d and each is yjd known as variable, feature, dimension or attribute.

- Hard partitioning gives result: C= {C1,.....,C_K} where (K≤N) and
  i. Ci ≠ ϕ, i=1, 2,...., N
  ii. $\cup_{i=1}^{K} Ci = Y$
  iii. $Ci \cap Cj = \emptyset, i, j = 1,2, \ldots . . K \; and \; i \neq j$
- Hierarchical clustering has different perspective of representing the output that is tree like structure, partition of Y,$P = \{P1, \ldots \ldots Pr\} \; where \; (r \leq N) \; and \; Ci \in P_{l \; and} \; Cj \in P_m$ and l > m imply $Ci \in Cj \; for \; all \; i, j \neq i, l \; m = 1,2, \ldots . r$

Clustering process:

Clustering process is step wise process which can be accomplished in the manner of their occurrences. So to discuss these steps in brief we will give them as follows:

- Feature selection or extraction: As pointed out by [1], feature selection is selecting distinguishing feature form set of candidates and extracting means

which it utilizes in the transformation to generate the useful and novel features from original ones as given by [2].

- Clustering algorithm design: In this step, constructing the criterion function and combining of selecting the algorithm according to the factors is done. Resulting clusters comes from the proximity measures taken. Every clustering algorithm is affected by measures. Next is to optimize the clustering solutions [3].
- Validation: After getting the results form cluster algorithm, the next step is to check the robustness of the system that whether it is up to the mark or not. These all can be checked by main three indices: External indices, internal indices, Relative indices.
- Result interpretation: Next step is to provide accuracy to user and provide a meaningful insight form original data so that efficient results can be provided.

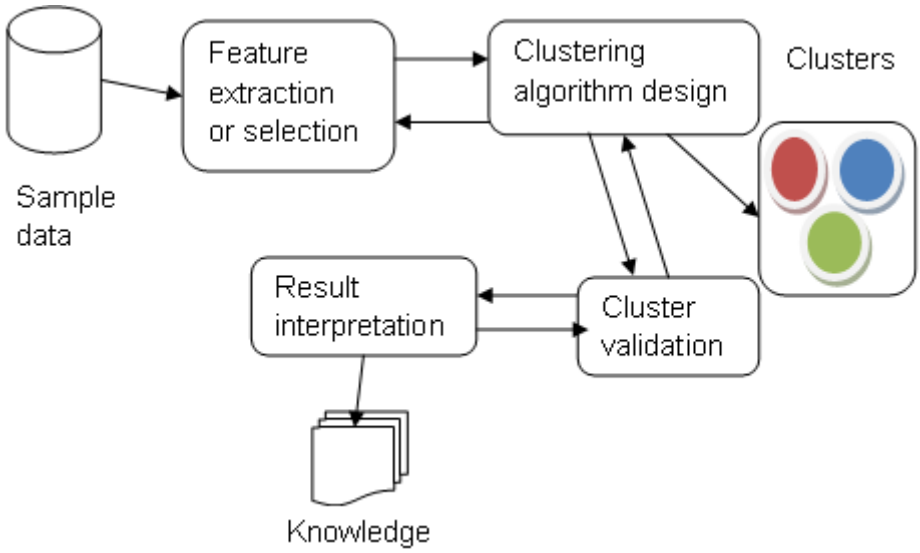


Figure 1.4 Clustering process

Clustering methods: There are various methods for clustering which act as a general strategy to solve the problem and to complete this, an instance of method is used called as algorithm.

Partitioning methods: This method simply partitions the dataset into n objects. K-partitions with n objects such that $k \leq n$. Different types of approaches are stated below:

- Grid based method uses grid data structure and at each step grid like structure is being followed [4].

- Subspace based uses subspace of actual document and it work with high dimensional data.
- Density based, its general concept is to increase the given cluster to cover the neighborhood exceeds some threshold value.

Relocation based methods have strategy on the conceptual point of view in which it identify the unknown parameters of the clusters [5,6].

There are number of algorithms being stated in this diagram. So it's very difficult to explain each and every algorithm at one place. So to give various features of algorithms so that it can compare the algorithm we state this in the form of table.

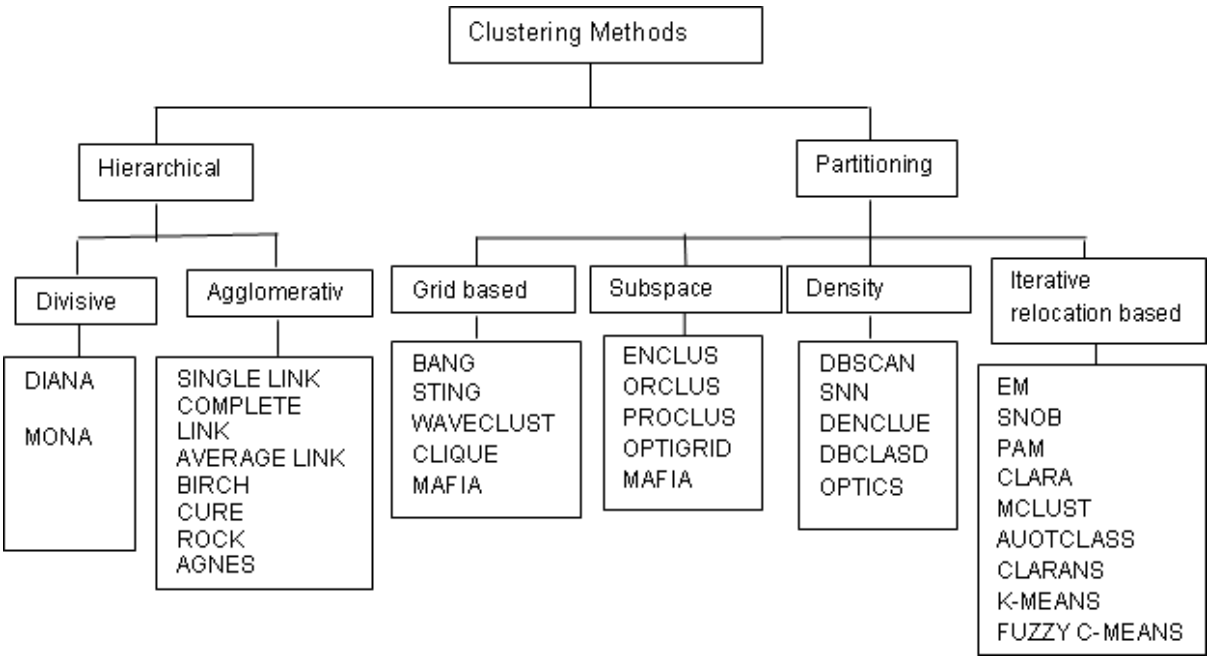
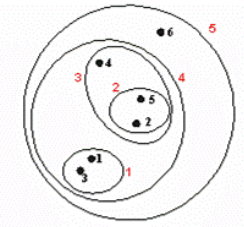


Figure 1.5 Clustering methods

Hierarchical methods: There is a tree like structure in this method. There are two approaches agglomerative and divisive [7].

- Agglomerative is bottom up approach which starts from leaf node having n clusters to reach up to root node. In intermediate steps it goes on merging each and every node. By this process the output will be the whole data set available.

- Divisive is the top down approach which starts with root node and by concentrating on each step it goes on dividing into number of clusters.



.

Figure 1.6 Agglomerative

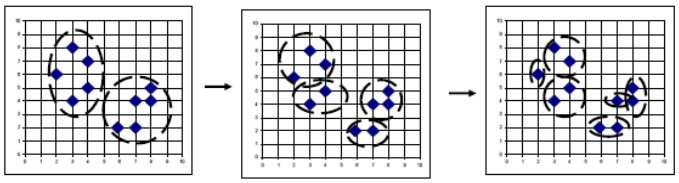


Figure 1.7 Divisive

We have different features on which the clustering algorithms can be explained for e.g. type of dataset they can handle, dimensionality of the dataset, size of the dataset, and the missing values it can handle.

Table 1.1: Different features of clustering algorithms [8]

Categories	**Algorithm**	**Type of dataset**	**Handling High dimensionality**	**Size of dataset**	**Handling noisy data**
Hierarchical	BIRCH CURE ROCK Chameleon	Numerical Numerical Categorical and numerical All type of data	No Yes No Yes	Large Large Large Large	No Yes No No
Partitioning	FCM K-means PAM CLARA CLARANS	Numerical Numerical Numerical Numerical Numerical	No No No No No	Large Large Small Large Large	No No No No No
Grid	OptiGrid CLIQUE STING Wave-cluster	Special data Numerical Special data Special data	Yes Yes No No	Large Large Large Large	Yes No Yes Yes
Iterative Relocation	EM SOMs COBWEB CLASSIT	Special data Multivariate Numerical Numerical	Yes Yes No No	Large Small Small Small	No No No No
Density	DBSCAN DBCLASD OPTICS DENCLUE	Numerical Numerical Numerical Numerical	No No No Yes	Large Large Large Large	No Yes Yes Yes

1.2.3 Regression:

Regression is another data mining technique which is based on supervised learning and is used to predict a continuous and numerical target. It predicts number, sales, profit, square footage, temperature or mortgage rates. All these can be predicted by using regression techniques. Regression starts with data set value already known. It is based on training process. It estimates the value by comparing already known and predicted values.

There are two types of regression techniques namely linear and non –linear.

- Linear regression: Linear regression is used where the relationship between target and predictor can be represented in straight line. Linear regression can be single predictor and multi predictor which has two or more predictors.

$$y = P1 \text{ x} + \text{P2} + \text{e}$$

Multivariate linear regression: The regression line cannot be visualized in two dimensional space.

$$y = P1 + P2\ x1 + P3\ x2 + \cdots + Pn\ xn - 1 + e$$

- Non- Linear Regression: In this case non linear relationship can be there and this cannot be represented as straight line. This can be represented as linear reaction by preprocessed data.

1.2.4. Association:

Association is another technique in data mining which is used to finding relationships. These are used to find patterns, correlations and associations. And also the casual structures between set of items or objects in relational databases, transactional databases and other repositories as well. There are some uncovered relationships which can be discovered by the association rules i.e. called as association rule mining. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns. They use the above mentioned criteria *support* and *confidence* to identify the most important relationships. *Support* is an indication of how frequent and regular basically the count

that items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true.

Apriori Algorithm:

This algorithm contains two main parts. Firstly it satisfies the minimum support which gives the frequent itemsets. Secondly the minimum confidence requires from the itemsets i.e. the minimum requirements of the itemsets.

ECLAT Algorithm:

Equivalent class transformation uses the set intersection theory and is the depth-first search algorithm. It suits for both sequential as well as parallel execution.

FP-growth Algorithm:

Frequent pattern growth algorithm states that there will be two passes in the first pass, the algorithm counts occurrence of items in the dataset, and stores them to 'header table'. In the second pass, it makes the FP-tree structure by inserting instances. There is condition that tem in each instance have to be sorted by descending order of the frequency in the dataset. It is required as that tree can be processed easily. It is not required that the items in each instance should meet minimum requirement of threshold. If many instances share most frequent items, FP-tree provides high compression close to tree root.

So to summarize all the techniques there is table which contains techniques, their algorithms and the application where we can use it for future.

Table 1.2: Techniques, Algorithms and Applications of data mining

Techniques	**Algorithms**	**Application**
Classification	Decision tree Naïve bayes Support vector machine Logistic regression	Rules, transparency Embedded application Text, narrow data Classical statistical technique
Clustering	Expectation-Maximization clustering (EM)	Fixed groups

	Hierarchical k-means Hierarchical O-cluster	Text mining, product grouping Gene and protein analysis
Regression	Linear regression Support vector machine	Classical statistical techniques Text and narrow data
Association rules	Apriori	Market basket analysis, best offer in future
Feature extraction	Single value decomposition Nonnegative matrix factorization	Feature reduction and text analysis
Attribute importance	Principal component analysis Minimum description length	Reduce noisy data, attribute reduction
Anomaly detection	One class Support vector machine	Unknown fraud cases or anomalies.

There are two ways in which data mining can be broadly used and have their research area. Data mining can be defined in the form of rules and it is also can be used to predict the future values.

Table 1.3: Summarization

Data Mining	
Descriptive (defining rules among data)	Clustering Association rules Sequence discovery Summarization
Predictive (predict the future value)	Prediction Classification Regression Time series analysis

Data mining can be used in various fields like predicting disease [9]. As stated in this paper there are different types of diseases predicted in data mining namely Hepatitis,

Lung Cancer, Liver disorder, analyze the Heart disease, Diabetes and Breast cancer disease predictions.

In educational system, data mining is very useful in every aspect. Thus, application of data mining in educational systems can be directed to support the specific needs of each of the participants in the educational process [10]**.** So with data mining techniques, the cycle is built in educational system which consists of forming hypotheses, testing and training, i.e. its utilization can be directed to the various acts of the educational process in accordance with specific needs. Some techniques for descriptive model include probabilistic models, clustering trees and association rules. Some examples of predictive techniques are: support vector machine, artificial neural networks, predictive rules and decision trees.

1.3 Relational data mining

Traditionally there is technique in data mining called as attribute value technique which requires data to be analysed and is stored in one single table. One single table means relation in which each instance is represented by fixed number of attributes. Data mining is basically analyzing the historical or previously stated data. Earlier our organizations were not so much wide i.e. there are small organizations and we have small amount of data which needs to be stored. But now a day's our organizations are growing at very high rate so the data is increasing as well. There is need to store whole data in our database to evaluate it in future. But in practical scenario data is stored in multiple, interconnected tables. To derive multi relational data to single value, there is loss of important information.

Therefore lot of research has focused on data mining techniques which learn directly from multiple tables in a database. These results in the enabling of discovering patterns which are not build from attribute based techniques. ILP means Inductive Logic Programming in which the area of research of relational table is being established. The relational data in ILP is represented as logic program which have an advantage that besides data they can also program knowledge. To induce proper knowledge, it provides facility to user in which they are allowed to express background information of problem domain.

Many algorithms has been proposed till now**.** We give the algorithm for finding frequent patterns from data streams with a case study and identify the research issues in handling data streams[10]. Data may be a sequence data, sequential data, time series, temporal, spatio-temporal, audio signal, video signal to name a few.

1.4 Ensemble methods

Most of the predictive data mining techniques used to generate one model that can be used to make predictions. On the other hand there is combinations of several models whose individual predictions are combined in different manners and this is called as Ensemble. Ensemble of model often outperforms their base models as stated by many researchers. Base model is the component models of the ensemble. This all happens if the base models perform reasonably well on novel examples and tend to make errors on different examples. Numerous techniques have been proposed for constructing ensembles from past years which result in an increased predictive performance, and hence, they have become very popular.

1.5 Motivation and contribution

Our work is mainly covering the area of predictive data mining more specifically in the area of classification. The main task to learn a predictive model is that it gives some sort of new values which are classified into predefined number of classes.

Over single classifier counterparts there is an abundance of classification ensemble methods exist which have clear advantage. They involve learning of set of classifiers, they are clearly less efficient and may be even more important, but less interpretable than just one single classifier.Despite the advantages of ensembles, these issues bring that they are much less often employed in problem domains where these matters are essential. Moreover, in high constrained learning environments where there is direct access to individual data examples is not available and only statistics is there, due to way of construction some popular ensemble methods just cannot be applied.

In this work, we propose some of the already applicable techniques aiming at making well established ensemble methods usable in various domains where they could clearly be beneficial. So we can summarize our contribution of the work being done:

- The **first contribution** lies in the relation of ILP. To upgrade the performance of predictive performance, ensemble methods give efforts in ILP. There is

drawback in this approach as the computational cost becomes so high. To solve this problem we update a first order decision tree algorithm to first order random forests inducer and note its behavior.

- The **second contribution** is in relation with comprehensibility of ensemble models. Main purpose to propose this model is to induce a single decision tree from an ensemble of decision trees. The new tree is derived by evaluating the heuristics of tests. Test is being conducted on the nodes from the class probabilities from the predicted results. We also provide an upgrade of the algorithm for ensembles.
- The **third contribution** is concerned only with learning from statistics. In this settings the statistics of data is provided not the individual data instances. We use statistics of data to evaluate the data.

1.6 Outline of the Thesis

The further organization of thesis is as follows:

Chapter 2: **Literature Review**

In this chapter various techniques to solve the standard problems are given. It also contains the survey of the tools, programming language and technique to be used.

Chapter 3: Research Problem

This chapter gives the problem statement which is going to be solved by the proposed approach.

Chapter 4: Proposed Solution

This chapter act as backbone to the research done. It contains the approach and algorithm to solve the problems.

Chapter 5: Testing and Results

In this chapter all the results have been shown in the form of tables and graphs.

Chapter 6: Conclusion and Future Scope

This chapter is devoted for the discussion with conclusion and extended to further research.

CHAPTER 2 LITERATURE REVIEW

In this chapter, work done by various researchers is summarized. It is divided into different sections.

2.1. Applications of data mining by using different techniques

Data mining techniques are used in many applications. Many users have designed prediction systems using these techniques. There is a study of various factors that affect academic performance and for that the data of pharmacy students have been taken focusing on which will help students to improve their performance[11]. Many factors have been included and on the basis of that the experiment has been done by Radaideh and Nagi[12]. A paper by Kriegel et al. focuses on building the classification model to predict the performance of employees[13]. By using classification techniques like decision tree, naïve bayes a prediction model is designed by Velmurugan[14]. Another paper by Sudha and Vijiyarani is on the prediction of diseases as heart diseases, diabetics etc. by using data mining techniques[8].

In a paper by Ngai et al. a review of the classification scheme for the application of financial fraud detection using data mining technique is done[15]. Use of K-means algorithm is very useful in designing many applications. Extension of K-means algorithm can be done to improve the performance as given by Huang[16]. A survey by André et al. shows different perspectives that in the data obtained by partitioning done by clustering ensembles, data can be improved by applying more steps and this all could be done through genetic programming approach[17].

As in unsupervised learning, there is no target attribute known in advance and there may be some time no comparison and correction in building groups. So to improve this, a new concept came into picture that is bounded rationality to reveal feature saliency in clustering problem designed by Aviad and Roy[18]. Comparison of various partition based clustering algorithms is done to distinguish among type of algorithms best suited for user's application by Sandeep et al.[19]. The new approach is being introduced for elder people living in old age homes to improve their way of living and to improve their health standards by Combes and Azema[20]. For validation of clusters different types of parameters are identified on the basis of which

clustering is done and relation between WB, Xu and Calinski- Harabasz index is stated [21].

Analysis of student performance can also be done by K-means algorithm where the predicting power of clustering algorithms and Euclidean distance for sum of squared errors, again academic data is taken and algorithms are applied[22,23].

Different types of prediction model for internet user are also proposed. Novel link prediction that is super edge prediction is being applied to create a super network model introduced by Liu et al.[24]. On large dataset the factors that affect performance can be taken care. So detailed study of this dataset is stated by Adhikari and Rao[25]. One of the paper works for histogram data by using Dynamic Clustering Algorithm with an automatic weighting step of the variables by using adaptive distances given by Irpino et al.[26]. This study is related to improve the shortcomings of csiFCM i.e. cluster size intensive fuzzy c mean algorithm. New method introduced is slibFCM i.e. cluster size insensitive integrity based FCM method by Lin et al.[27].

A paper by Xiao and Fan focuses on analyzing the large data in BAS building automation system and also improve the building operational performance[28]. For multivariate functional data the new model based clustering algorithms is proposed by Jacques and Preda [29]. Using hybrid clustering approach, mining of categorical sequences from data can be done by as suggested by Angelis and Dias [30].

2.2 Tools

Various tools have been used for data mining as discussed here:

(i) Tool 1-Orange

Orange is an Open source data visualization and analysis tool. It was developed at Bioinformatics laboratory at the faculty of computer and information science. Orange has some features of visualization such as scatter plots, bar charts and trees to dendrograms, network and heatmaps. This is basically devoted to machine learning methods for classification. Classification uses two types of objects: learners and classifiers. Given the data, they return the models that can predict the outcome for any data instance [31].

(ii) Tool 2- WEKA

WEKA stands for Waikato Environment for Knowledge Analysis and is developed in Java programming language. It is not capable for multi relational data mining. Pre-processing tools are known as filters and these are available for discretization, normalization, resampling, attribute selection, transforming and combining attributes. One additional feature is that data sources, classifiers etc are called as beans and these can be connected graphically [32].

(iii) Tool 3-SCaVis

Scientific **C**omputation and **Vis**ualization Environment

It provides environment for scientific computation, data analysis and data visualization designed for scientists, engineers and students. It provides freedom to choose a programming language, freedom to choose an operating system and freedom to share code.The program incorporates many open source software packages into a coherent interface using the concept of dynamic scripting. There is provision of multiple clipboards, multi-document support and multiple Eclipse-like bookmarks Extensive LaTeX support: a structure viewer, a build-in Bibtex manager, LaTeX equation editor and LatexTools[33].

(iv) Tool 4- Apache Mahout

A tool for clustering. Clustering can be done which includes various algorithms: Canopy Clustering, k-Means Clustering, Fuzzy k-Means, Streaming k-Means, Spectral Clustering by Sean Owen and Sebastian Schelter[34]. Its goal is to build scalable machine learning library.

(v) Tool 5- R Software Environment

It provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, classification, clustering as given by Baker and Yacef[35]. It is a free software environment for statistical computing and graphics.

(vi) Tool 6- ML Flex

ML uses machine learning algorithms to derive models from independent variables with the purpose of predicting the values of a dependent (class) variable.

(vii) Tool 7- Databionic ESOM (Emergent Self –Organizing Maps) tool

Preprocessing, Training, Visualization, Data analysis, Clustering, Projection, Classification these all can be performed using this tool. The process of SOM training adapts the grid of prototype vectors to the given data creating a 2-dimensional projection that preserves the topology of the data space. Online training, there is immediately update of best match but in batch training all the best matches are being collected and then update if performed collectively [36]. The process of SOM training adapts the grid of prototype vectors to the given data creating a 2-dimensional projection that preserves the topology of the data space.

(viii) Tool 8-NLTK (Natural Language Tool Kit)

NLTK is a leading platform for building Python programs to work with human language data. As stated by Sandeep et al. NLTK defines several classifier classes: Conditional Exponential Classifier, DecisionTree Classifier, Maxent Classifier, NaiveBayes Classifier, Weka Classifier.It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and an active discussion forum.

(ix) Tool 9-ELKI

Environment for **D**eveloping **KDD**-Applications Supported by **I**ndex- Structures

The focus of ELKI is research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection. It is open source data mining software written in Java. In ELKI, data mining algorithms and data management tasks are separated and allow for an independent evaluation. The fundamental approach is the independence of file parsers or database connections, data types, distances, distance functions, and data mining algorithms[37].

(x) Tool 10-UIMA

Unstructured Information Management Architecture

It enables application to be decomposed into components. Working of framework is to manage these components and flow between them. Basic availability is frameworks, components and infrastructure [38,39]. In this basically there is analyzing unstructured data such as video, audio and text. Large amount of unstructured information is being analyzed so that we can get relevant information.

(xi) Tool 11-GraphLab

One main advantage of this is that we can implement our own algorithm on top of our graph programming API[40]. GraphLab has its basic features that it has several algorithm already implemented in its toolkit.

(xii) Tool 12-MLPY

Machine **L**earning **P**ython

Cluster analysis can also be done for dimensionally reduction and wavelet transformation by this tool. Various different algorithms are also there like feature ranking, resampling algorithm, peak finding algorithm, error evaluation.

(xiii) Tool 13-KEEL

Knowledge Extraction Evolutionary Learning

KEEL is open source java software which have license of GPLv3. GPLv3 means General Public License which has latest version 3. It allows users to have the access of behavior of evolutionary learning and basic soft computing based techniques for various kinds of data mining problems to be handled.

(xiv) Tool 14-Scikit-learn

Scikit- learn uses the matplotlib package for plotting charts. Scikit-learn is a free package in Python that extends the functionality of NumPy and SciPy packages with numerous DM algorithms. The package supports most of the core DM algorithms. However, several significant DM algorithm groups have been omitted currently, including classification rules and association rules.

Table 2.1 General introduction of tools

Tool	Developed	Main language supported	Aim
Orange	06-2004	Python	Visual data analysis
WEKA	1997	Java	General ML package
Kernlab	04-2004	R	Kernel based classification/ Dimensionality reduction
Dlib	2006	C++	Portability, correctness
Nieme	09-2006	C++	Linear regression, Ranking, Classification
Java-ML	08-2008	Java	Feature selection
pyML	08-2004	C++, python	Kernel methods
Shogun	1999	C++	General Purpose ML Package with particular focus on large scale learning; Kernel Methods; Interfaces to various languages
Mlpy	02-2008	Python	Basic algorithms
Torch7	01-2002	C++	Neural networks
Pybrain	10-2008	Python	Reinforcement learning
Scikit-learn	2007	Python	General Purpose with simple API and numpy / scipy idioms

Table 2.2 Comparison of various tools on the basis of operating system supported:

Tools	Linux	Windows	Mac OSX	Other Unix
Orange	Yes	Yes	Yes	Yes
WEKA	Yes	Yes	Yes	Yes
Kernlab	Yes	Yes	Yes	Yes
Dlib	Yes	Yes	Yes	Yes
Nieme	Yes	Yes	Yes	Yes
Java-ML	Yes	Yes	Yes	Yes
pyML	Yes	No	Yes	No

Shogun	Yes	Yes	Yes	Yes
Mlpy	Yes	Yes	Yes	Yes
Torch7	Yes	Yes	Yes	Yes
pybrain	Yes	Yes	No	No
Scikit-learn	Yes	Yes	Yes	Yes

Table 2.3 Comparison on the basis of language bindings:

Tools	**Python**	**R**	**Matlab**	**Octave**	**C/C++**	**Command line**	**Java**	**C#**	**Lua**	**Ruby**
Orange	Y	N	N	N	N	N	N	N	N	N
WEKA	N	N	N	N	N	N	Y	N	N	N
Kernlab	N	Y	N	N	N	Y	N	N	N	N
Dlib	N	N	N	N	Y	N	N	N	N	N
Nieme	Y	N	N	N	Y	N	Y	N	N	N
Java-ML	N	N	N	N	N	N	Y	N	N	N
pyML	Y	N	N	N	N	N	N	N	N	N
Shogun	Y	Y	Y	Y	Y	Y	Y	N	N	N
Mlpy	Y	N	N	N	N	Y	N	N	N	N
Torch7	N	N	N	N	Y	Y	N	N	N	N
Pybrain	Y	N	N	N	N	Y	N	N	N	N
Scikit-learn	Y	N	N	N	N	N	N	N	N	N

Table 2.4 Comparison on the basis of general features:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
GUI	Y	Y	N	Y	Y	N	N	N	N	Y	Y	N
One Class Classification	N	Y	Y	Y	N	N	Y	Y	N	N	N	Y
Classification	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Multiclass classification	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
Regression	Y	Y	Y	Y	Y	N	Y	Y	N	Y	Y	Y
Structured Output Learning	N	N	N	N	Y	N	N	Y	N	N	N	N

Pre-Processing	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y
Built-in Model Selection Strategies	Y	Y	Y	Y	N	Y	Y	Y	N	N	N	Y
Visualization	Y	Y	N	Y	Y	N	Y	N	Y	Y	Y	Y
Test Framework	-	Y	N	Y	Y	Y	N	Y	N	N	N	Y
Large scale learning	N	N	N	Y	Y	N	N	Y	Y	N	N	N
Semi-supervised Learning	N	N	N	N	N	N	N	N	N	N	N	-
Multitask Learning	N	N	N	N	N	N	N	Y	N	N	N	N
Domain Adaptation	N	N	N	N	N	N	N	Y	N	N	N	N
Serialization	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Parallelized Code	N	Y	N	Y	N	N	N	Y	N	N	N	Y
Performance Measures	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y

1.Orange, 2. Weka, 3. Kernlab, 4. Dlib, 5. Nieme, 6. Java-ML, 7. PYml, 8. Shogun, 9. MLpy, 10. Torch , 11. PY-brain, 12. Scikit learn

Table 2.5: Comparison of tools on the basis of file formats supported

Tools	**Binary**	**Arff**	**HDF5**	**CSV**	**Excel**	**Protobuf**	**SVM light format**
Orange	N	N	N	Y	Y	Y	N
WEKA	Y	Y	N	Y	N	N	Y
Kernlab	N	N	Y	Y	Y	N	N
Dlib	N	N	N	N	N	N	Y
Nieme	N	N	N	N	N	N	Y
Java-ML	N	Y	N	Y	N	N	N
pyML	N	N	N	Y	N	N	Y
Shogun	Y	N	Y	Y	N	N	Y
Mlpy	N	N	N	Y	N	N	N
Torch7	N	N	N	Y	N	N	N
Pybrain	Y	N	N	N	N	N	Y
Scikit-learn	Y	N	N	Y	N	N	Y

So after discussing the various factors of different tools there is summarization of the tools used for various techniques of data mining. Below is the figure which specifies that which tool is used for which technique of data mining.

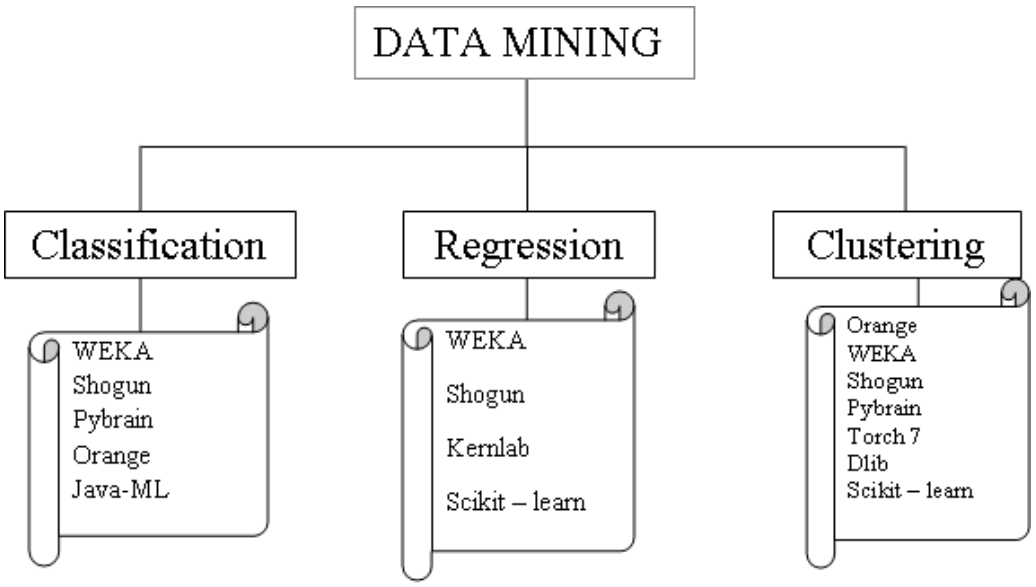


Figure 2.1: Tools and Data Mining Algorithms

2.3 Decision tree learning

One of the most popular and widely used methods in data mining is Decision tree learning. They exhibit some useful properties. They have a well-inferred theoretical basis and also the induction algorithms have a low computational complexity. These are relatively robust to noisy data and handle it accordingly. As decision trees are also used within this text, we will elaborate on decision trees in this section. First of all we describe how a decision tree is represented and how it can be used for prediction. Then an algorithm is given that shows a decision tree in a top-down manner.

Representation of decision tree:

Predictive model i.e. decision tree have a phenomena to map observation of an item to the conclusions of the item's target value. This target can be class label or numerical value. As we have different characteristics of techniques of data mining. In class label values the tree models are called as classification trees and in numerical values the tree models build by the regression techniques. In tree structures we have leaf nodes

and parent nodes. Leaf node represents the prediction and branches represent the conjunctions of features lead to prediction.

A decision tree for a domain X is defined recursively as follows:

• $leaf$ (C) is a decision tree for X, where C is a class value in a numerical value in a regression tree or a classification tree.

• int_node (t, S) is a decision tree for X if and only if t is a function from X to some set R. t is a test and S is a set of pairs such that $\forall\, r \in R$ so $(r, tr) \in S$ and tr is a decision tree for X.

i.e., each internal node in a decision tree contains a test t and associates a subtree tr with each possible outcome r of the test t .

There is conjunction of attribute tests, which is represented by each path from the root of the tree to a leaf and the tree itself to a disjunction of these conjunctions.

Prediction instance e can be received by sorting it down the tree, starting at the root until it reaches the last node i.e. leaf node. There is internal node n of the tree and the test tn is applied to the instance. Depending upon the outcome of the instances, tree is further sorted down the subtree tr. This loop will repeat until the leaf node is achieved the final prediction is shown on the leaf node.

Top down induction of decision tree:

It's being stated by (Zantema and Bodlaender 2000) that the task of finding the smallest tree that fits a certain data set is NP-hard[40]. And the decision tree induction algorithms usually use heuristic, greedy search techniques to build the trees. Most of the algorithms that learn decision trees construct the tree starting from the root to the leaves. This method is referred to as Top Down Induction of Decision Trees (TDIDT)[41]. Top down induction can be mostly implemented by systems like CART [42], C4.5 [43] and the reimplementation of the C4.5 i.e. J48 [44].

The main function of this method is that it recursively chooses a test that splits the data into maximally homogeneous subsets and re-applies this procedure on the obtained subsets until a stopping criterion. Stopping criteria is met only if its all subsets are homogenous at its middle stage and at lower bounds. At the end the nodes reach at the leaf node. The local model predicted by the leaf depends on the data mining task at hand. For a regression tree it predicts the mean of their target values and in a classification tree, ending up in that leaf which have the most frequent class among training examples.

Following is the basic algorithm for the top down induction that it works for the decision tree.

Algorithm 2.1 A generic TDIDT algorithm

$procedure\ TDIDT\ (E: examples)$:

$P\_T := \ set\ of\ all\ possible\ tests$
$T := \ argmax\ r \epsilon P\_T\ quality(P\_T, E)$
$If\ Stop\ (T, E)\ then$
$\ Return\ leaf(local\_model(E))$
$Else$
$\ P\ = partition\ induced\ on\ E\ by\ T$
$\ For\ all\ Pi\ in\ P\ do$
$\ ti := \ TDIDIT\ (Pi)$
$\ end\ for$
$\ return\ int\_node(T, Ui\ \{(i, ti)\})$
$end\ if$

There will be two possibilities of the stopping criteria one is either strong or weak. When the stopping criterion is too strong, the decision tree might not be able to capture the knowledge available in the data. A way to deal with the difficulty of estimating precisely when to stop growing a tree that we should allow the oversized tree to grow and then after it grows prune away those branches that do not seem useful. Several methods for post-pruning are in use.

Now next condition is when the stopping criterion is too weak, decision trees tend to overfit. This is certainly in the presence of noise in the data. This means that it does not generalize well to unseen data, although the tree performs well on the training data. In this case there might exist another decision tree that has a higher predictive performance on the test data but perform worse on the training data.

Quality of a test:

Next important step in this recursive tree induction process is the choice of the test that main function is to be put in the node. The quality of a test is usually evaluated using a heuristic function. This function measures how well a test can partition the data into homogeneous data with respect to the target attribute subsets. Main goal is that it selects a test that reduces the impurity of the induced subset. For regression, heuristics are used that reduce the variance in the subsets[42]. On the other hand, for

classification often used heuristics are information gain [43]**,** gain ratio [43] or gini-index [42].

The little more details about information gain. The information gain of a test T with respect to a data set E is defined in terms of class entropy as follows:

$$\mathrm{IG(t,E)} \;=\; \mathrm{entropy(E)} \;-\; \sum_{Ei \in p} \frac{|\mathrm{Ei}|}{|\mathrm{E}|} \mathrm{entropy(Ei)}$$

where P is the partition on E induced by t and entropy is defined as follows

$$\mathrm{entropy(E)} \;=\; -\sum_{i=0}^{c} \mathrm{p(c_i, E)} \log_2 \mathrm{p(c_i, E)}$$

where c is the number of classes, the ci are the classes and $p(ci, E)$ is the proportion of examples in E that belong to class ci

2.4 Ensemble

One of the earliest works on ensemble systems is being done by Dasarathy and Sheela's 1979 paper, which discusses partitioning the feature space using two or more classifiers[45]. Since these works, research in ensemble systems has expanded rapidly. These will be appearing often in the literature under many creative names and ideas. The long list includes stacked generalization [46], change-glasses approach to classifier selection [47], dynamic classifier selection [48], composite classifier systems [45], mixture of experts [49,50], combination of multiple classifiers [51-54], classifier fusion [55-57], pandemonium system of reflective agents [58], and classifier ensembles [59, 60] among many others.

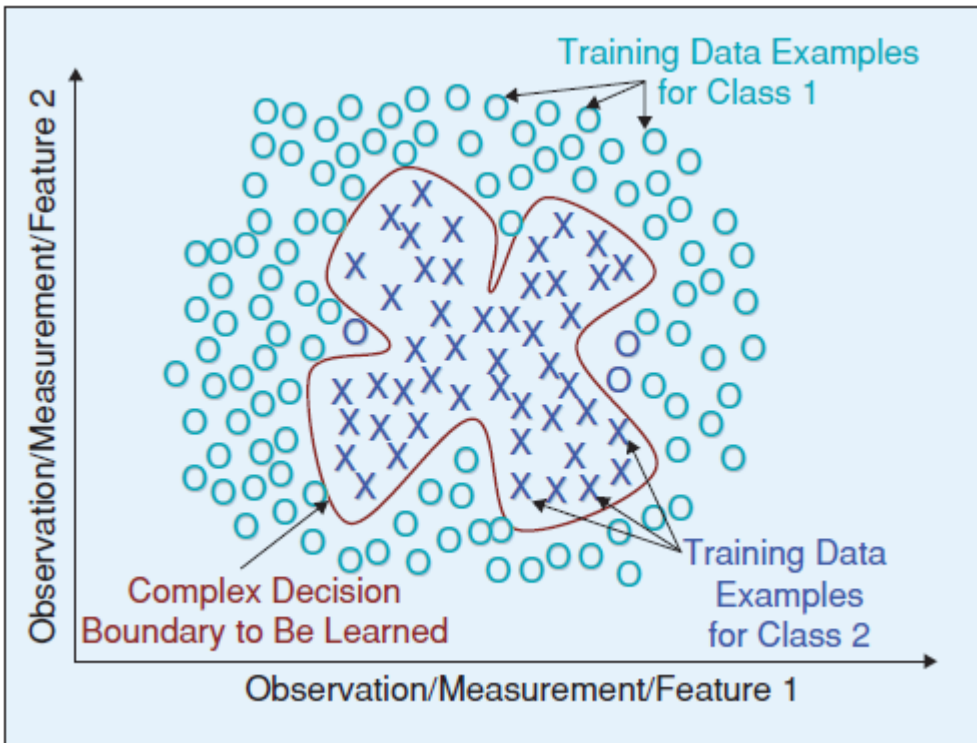


Figure 2.2: Complex decision boundary that is not possible to learned by linear or circular classifiers [61]

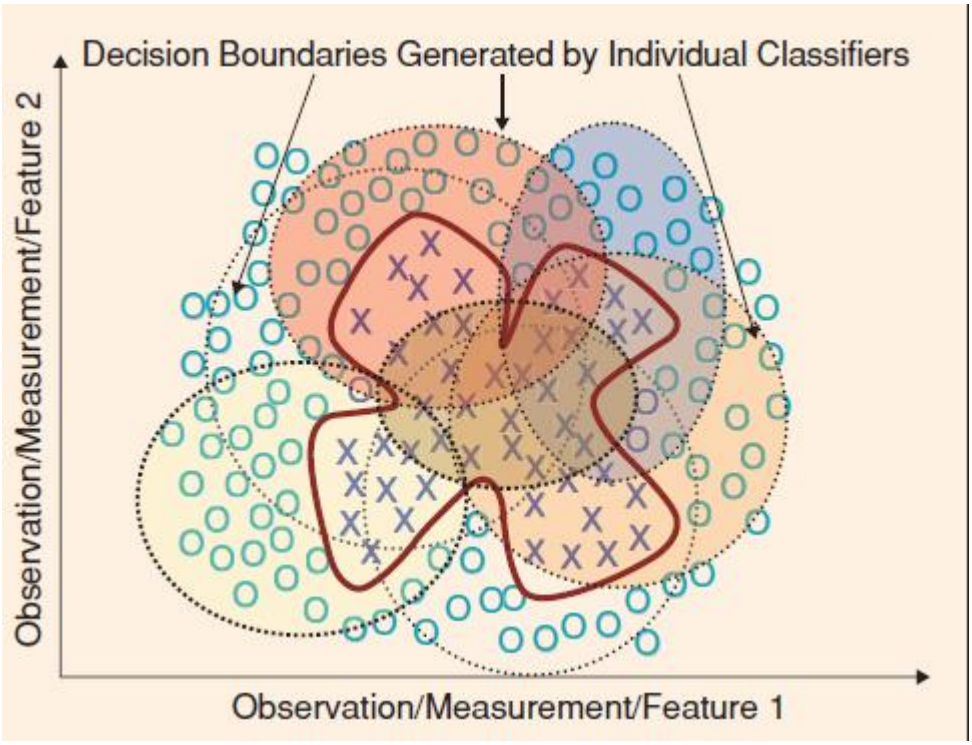


Figure 2.3: Ensemble of classifiers spanning the decision space [61]

The archetypes of these approaches differ from each other with respect to the specific procedure basically used for generating individual classifiers, and/or the strategy carried for combining the classifiers. There are specifically two types of combination: Classifier selection and Classifier fusion [61].

- In *classifier selection*, each classifier is trained so that it becomes an expert in some local area of the total feature space. The combination of these selected classifiers is then based on the given feature vector. The classifier selected with feature vector means trained with data closest to the vicinity of the feature vector, in some distance metric sense and is given the highest credit. One or more local experts can be nominated to make the decision [62–64].
- In *classifier fusion*, all classifiers are trained over the entire feature space. This step involves, the classifier combination process begins with merging the individual (weaker) classifier designs to obtain a single (stronger) expert of superior performance.

For this approach we have some examples which include bagging predictors [65], boosting [66], [67] and its many variations. The combination may apply to the class specific continuous valued outputs of the individual or to classification labels experts [68, 69].

A number of authors have given theoretical analysis of various strategies which are commonly used in multiple expert fusion: for example, theoretical models were developed in [70,71] and in this they are used for combining discriminant functions. In this six commonly used combination rules were compared for their ability to predict posterior probabilities in [72]; and for multiple expert fusion a Bayesian theoretical framework was developed. In this framework to estimation errors the sensitivity of various combination schemes was analyzed to propose a plausible model that explains such behaviour. There is an immense literature on classifier combination which can be found in Kuncheva's recent book and this is the first text devoted to theory and implementation of ensemble based classifiers and references therein. The field has been developing so rapidly that an international workshop on multiple classifier systems (MCS) has been established, and the most current developments can be found in its proceedings.

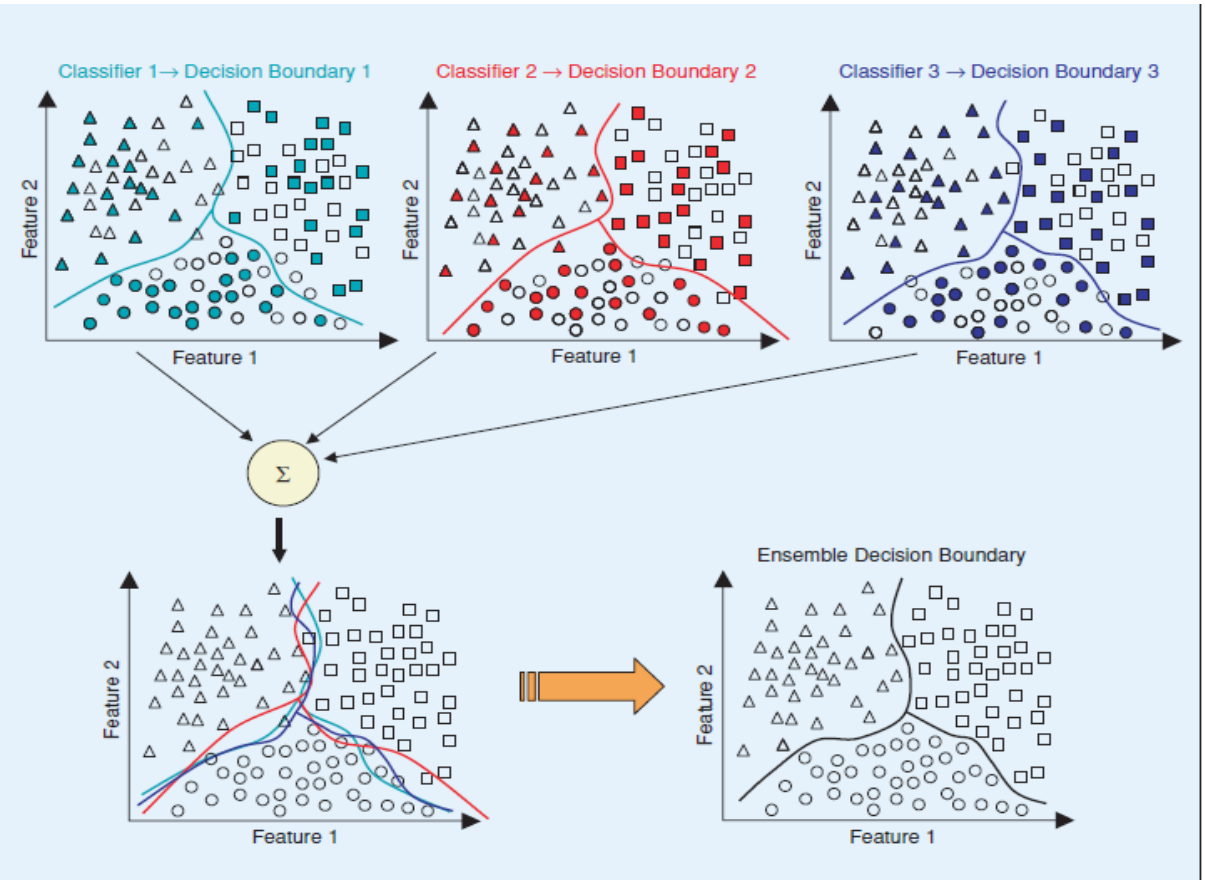


Figure 2.4: Combining classifiers that are trained on different subsets of the training data [61]

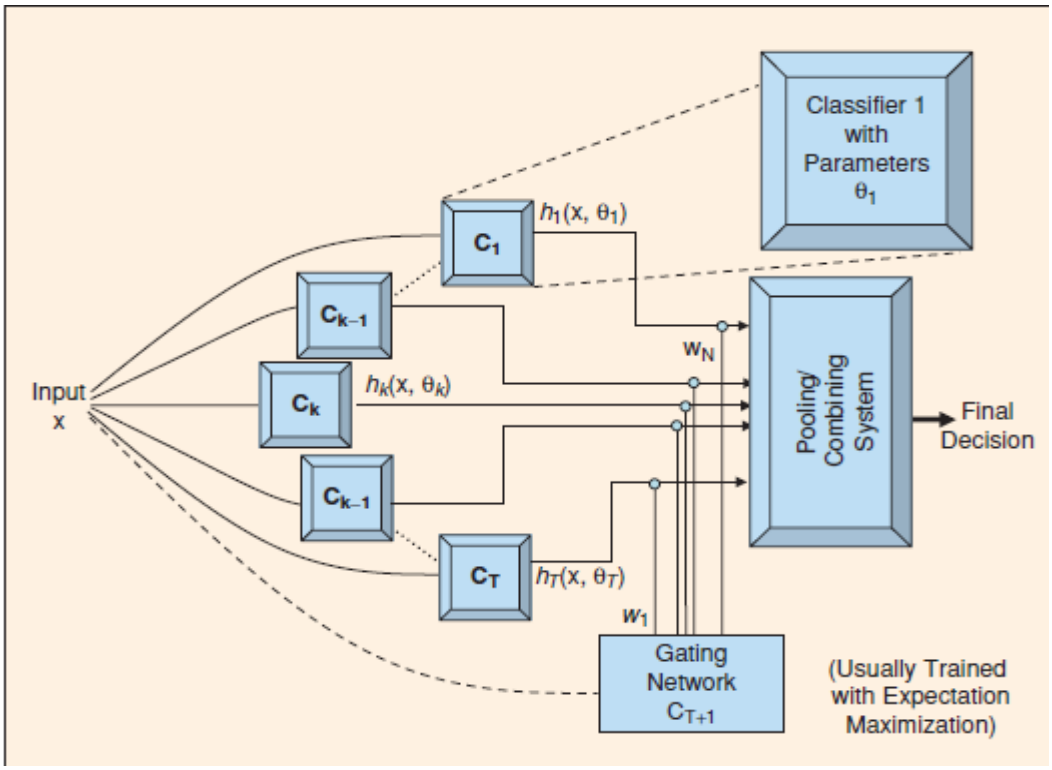


Figure 2.5: Mixture of experts [61]

2.5 R programming language:

R programming language provides the environment for Data Analysis and Graphics.
R is an integrated suite of software which give us facilities for calculation, data manipulation, and graphical display. Among other things it has:

- a suite of operators for calculations on arrays, in particular matrices.
- a simple, well developed and effective programming language (called 'S') which have conditionals, user defined recursive functions, loops, input and output facilities. Indeed most of the system supplied functions are themselves written in the S language.
- an effective storage facility and data handling.
- a large, coherent, integrated collection of intermediate tools for data analysis.
- graphical facilities for data analysis and display either directly at the computer or on hardcopy.

The term "environment" is intended to characterize it as a coherent system which is fully planned, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.
After seeing its features there are various reason that why to learn this language. Some of them are as follows:

- It is an open source language
- It relates to other languages as well
- It is an advanced statistical language
- It has wonderful concept of visualization
- It also supports cross platform
- It supports extension as well
- It is extremely comprehensive
- It has vast community
- As it supports cross platform so it is flexible
- R has access to powerful analytics
- With R we collaborate with various applications. Also it provides facility to connect to MySQL, Apache and it provides Mash-ups as well.

CHAPTER 3 RESEARCH PROBLEM

3.1 Barriers in the previous work

With the growth in usage of internet now-a-days, data is growing exponentially. This growing data needs to be stored in databases. This has to be stored in order to support scalability and store large amount of data efficiently. This data can be stored and used for future analysis so that useful results can be derived from it. Basically data mining is collection of techniques which are used for efficient discovery of the previously useful, understandable, novel, unknown and valid patterns in the database. These patterns have capability to make useful decision. Data mining is complex process which involves various critical steps. Now one of the technique classification is being used by researchers for their research in the area of the prediction systems.

Classification is the technique of classifying the data or grouping the data into various classes by analyzing various factors. It is a supervised learning technique in which output label is already given. There are numbers of objects which are known. We have to predict one object from those. The object which is to be predicted is the dependent attribute or the output attribute and the rest attributes are the independent attributes or the input attributes. Prediction system is the system which gives the occurrence of future events in advance. The prediction system is also known as decision support system. This system has more importance if it gives more and more accurate results. The accuracy of system is one of the key factors in many applications. Mostly in medical field and industrial field this system plays a vital role. If we implement one model and classify our system, then there we get results mostly with higher accuracy. But the situation comes in which our system does not respond to the test data we give to the system. So to avoid this worst case scenario the concept of ensemble modelling comes into picture.

An Ensemble of classifiers is collection of n classifiers whose individual predictions are combined in different manner.

3.2 Problem statement

With the ability of combining the classifier models i.e. ensemble modelling we can improve the accuracy of prediction system. There are several mathematically sound reasons for considering ensemble systems, but the intrinsic connection to our daily life experiences provide an undeniably strong psychological pretext: we use them all

the time. Seeking additional opinions before making a decision is an innate behaviour for most of us, particularly if the decision has important financial, medical or social consequences. To make our system respond in worst conditions we use mixture of expert ideas and make it robust as if one model fails for some data value then other will respond to it.

There are various reasons why we go for this method. In some cases, we have problem of generalized problem (mostly in the case of neural network). So combining the outputs of several classifiers by averaging may reduce the risk of an unfortunate selection of a poorly performing classifier. The averaging may or may not perform as the best classifier in the ensemble, but it certainly reduces the overall risk of making a particularly poor selection. In some situations there is large volume of data. So to train a classifier with such a vast amount of data is usually not practical; partitioning the data into smaller subsets, training different classifiers with different partitions of data, and combining their outputs using a combination rule often proves to be a more efficient approach. There is a condition when there is small amount of data. So there we need adequate amount of training. When we have heterogeneous data collected from various sources then one classifier will not handle that data. So we need ensemble.

3.3 Objectives

In the light of above discussed research gaps following objectives have been formulated.

- There should be an algorithm which has ability to check the realities of noise, outliers and overlapping data distributions, however, make such a classifier an impossible proposition. At best, we can hope for classifiers that correctly classify the field data *most of the time*.
- To design an approach that should handle the noisy data outliers and predict with high accuracy.
- The proposed model should predict the class labels, check their validity and give maximum accuracy.
- The proposed model should have the facility to handle the worst case scenario where we have data that is unpredictable by one model.
- This model can be the amalgam of more than one models to give accurate results.

3.4 Research Methodology

Ensemble of models of classifier is the main area of focus. There are different models available for classification. These models are applied on various dataset and results are noted. The above mentioned experiment will be done in R programming language. There is use of R language and RStudio which is an IDE (Integrated environment development). RStudio is the frontend for R language. It has many benefits over R as there are already installed packages.

There is availability of various algorithms to solve the problem. For solving any problem and getting desired results from it, we can apply some common procedure on the dataset. There are different combination rules to combine the models. We can combine the models by the methods as by:

- Majority voting :- unanimos,simple,plurality
- Modified dempster with prior based
- Dubois-prade
- Yager
- Unweighted voting
- Modified dempster with uniform prior
- Disjunction
- Dempster-shafer combination
- Algebric connectives :- maximum , minimum,mean,median and product
- Fuzzy integral
- Decision templates
- Discounted dempster shafer combination
- Borda count
- Mixture of logical experts(mle)
- Hierarchical mle
- Associative switch
- Class set reduction
- Stacking
- Behavior Knowledge Space (BKS)

So we can apply any combination rule for our model to make ensemble model. Mostly the algebraic connectives are applied but it depends upon the programmer.

CHAPTER 4 PROPOSED SOLUTION

4.1 Proposed Work

The proposed approach is amalgamation of the methods of classifiers with that of genetic algorithm to bring out a hybridized ensemble based system which provides robust platform to classify the class labels to make a model and this model and new data combined is used for prediction for the future events. The work has been performed on two datasets. One of the dataset is being taken from TUNEDIT- DASL (Data and Story Library) and second is the dataset collected by the questionnaire.

Now the proposed algorithm will work on dataset given to it. This algorithm is divided into steps which are as follows:

ALGORITHM 4.1

Step1: Input the file which has description of the data.

Phase 1:

Step 2: Preparing data for the current phase by analyzing, pre-processing and cleansing.

Step 3: After *step 2* we have the data which is to be given to machine.

Step 4: Apply various models on this dataset. To exercise the data with working to make results.

Step 5: To calculate importance of attributes and ranked according to their significance in evaluating the results.

Step 6: Decision rules are built.

Phase 2:

Step 7: Calculate results by generalizing functions.

Step 8: Next step is to ensemble the models by applying combination rules.

Step 9:

If: accuracy of the ensemble models is greater than the each model

$$Acc_{en} > Acc$$

Write results to file (write.csv)

Else: Drop that models and apply different models and go back to *phase 1*

Step 10: Output the final results into one file.

In the proposed algorithm, in Step 1 we pass the description file to the system and the input file which contain data. In this the system read the file by passing the command to the system. There are two phases of the algorithm which contains different steps having different functioning. In phase 1 step 2, pre-processing on data is applied so that noisy data can be removed and data is passed to system. Now in step 3 various models have been applied to our datasets. We have taken two datasets for experiment as explained earlier. We applied various models on that but the successful models for these datasets are *rpart, rf, glm, ksvm*. These all models have different functioning and way to do evaluation of results. The importance of every feature is being calculated and the entropy is calculated to improve performance. After this we have to build the decision rules so that it can be implemented in diagrammatic way. After that the functions have been calculated by generalizing it.

Step 8 is the most important step for this algorithm to complete. In this step the combination rule is applied to the results formed by previous step. There are different types of combination rule available to amalgamate the results. In proposed algorithm we applied majority voting technique to obtain results.

(i) on which all classifiers agree (*unanimous voting*);

(ii) predicted by at least one more than half the number of classifiers (*simple majority*); or

(iii) that receives the highest number of votes, whether or not the sum of those votes exceeds 50% (*plurality voting* or just *majority voting*).

In voting we have the concept:

$$\sum_{k=1}^{n} \mathrm{d_{k,n}} = max \sum_{k=1}^{n} \mathrm{d_{k,n}}$$

And the accuracy of the models is being calculated. There is threshold accuracy for the system i.e. set according to the accuracy of individual model. If the ensemble

accuracy is greater than this accuracy then these models are acceptable. This states that our system is performing well and it is improved. But in case if it fails or accuracy is not upto the mark then this model is not to be used for the system to be trained. After completing all steps the output is saved into the file by the command.

There is option of using various models but the models used for our system are explained as below:

1) **Recursive Partitioning and Regression Trees: "rpart":** The rpart programs build regression models or classification of a very general structure. It has a two stage procedure; the resulting models can be represented as binary trees. First the single variable is found which best splits the data into two groups. The second stage consists of using cross-validation part to trim back the full tree.
2) **Random Forest: "rf":** Random Forest is useful for growing many classification trees. It is used to classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives the classification phenomena, and the tree "votes" for that class. It efficiently runs for large datasets and unexcelled in accuracy among current algorithms. It works well with missing data as well. Capability of random forest can be extended to unlabelled data as well. Experimental results of random forest are used for variable interactions as well.
3) **Generalized Linear Model: "glm":** When previously experiments were carried out they worked with regression model where response variables are quantitative and normally distributed. So there we need to take attention to response variables which are discrete and error terms will not go to normal distribution. So glm comes into picture. In this mean is dependent on the explanatory variables through link function.
4) **Kernlab Support Vector Machine: "ksvm":** Kernel based methods are used for machine learning methods such as classification, regression, novelty detection etc. Kernlab supports Support Vector Machine technique. Ksvm use one-against-one approach for multi class classification and uses k(k-1)/2 for binary classifiers. One-against-one approach is also called as pairwise coupling.

4.2 Proposed model:

The proposed model for any type of problem to be solved is the basic model which is applied to solve it. There is the division of each step which includes the different functioning to be executed. So the flowchart for the various steps is as follows:

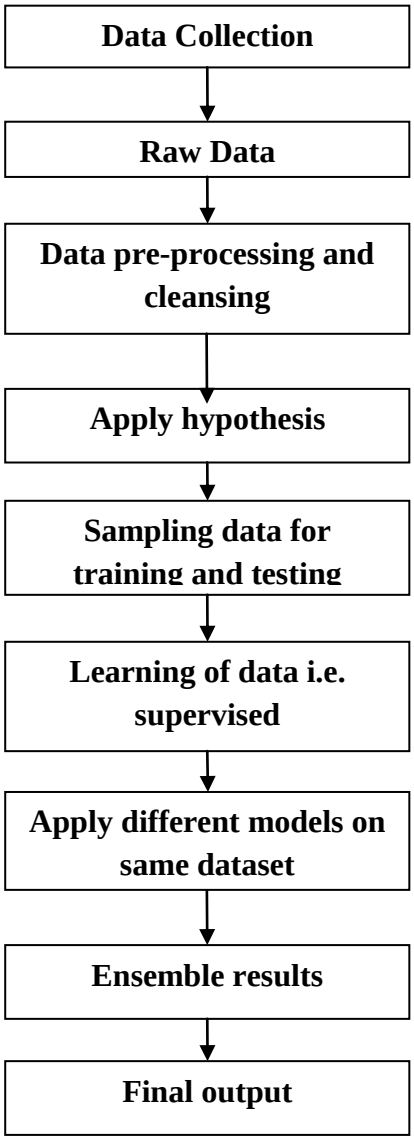


Figure 4.1: Proposed Model

Model: $f(x) = a_0 + \sum_{n=1}^{N}(a_n \mathrm{T_n}(\mathrm{x}))$

- $Tn(x)$: basis functions or base learners
- i.e., linear model in a very high dimensional space of derived variables

Learner characterization: $T_n = T(x: p_n)$

- p_n: a specific set of joint parameter values e.g., split definitions at internal nodes and predictions at terminal nodes
- $T(x; p_n)$: function class i.e., set of all base learners of specified family

CHAPTER 5 TESTING AND RESULTS

Testing and Results

The experiment is carried on the TUNEDIT-DASL and the data collected from students by questionnaire. The first data is about the company data which consists of various features as their asset values, their sales etc. to predict the sector in which company categorized. For the prediction we apply different models to the system from which we are successful with the three models. The models are then combined so that the accuracy of system is increased in order to get more efficient results. The coding is done in R programming language.

In the given experiments we have the features which are used for classification of the class. Classification is done by applying different classifier models and blending these models by taking two or models at a time. Apply the combination rules to the blend and keep on doing this until the system gives the better results.

The following models are implemented for the datasets to obtain the desired results:

Table 5.1: Methods used for system

Model Name	**Method**	**Package**
Decision Tree	rpart	rpart
Random Forest	random forest	rf
Support Vector Machine	kernlab	ksvm
Linear Model	generalized model	glm

Now the experiment is done for two datasets having partition 70-30% that means it has 70% training data and the 30% testing data. The consistency of the system is checked by the cross validation. The k-fold means that in this experiment there is 10-fold validation is being on different seed values. Experiment is being done on different seed values. Seed values are the random number generators which are used to generate the random number that will be taken for testing of data. There will be different seed values are used for one experiment and it is mentioned by programmer. Table 5.2-5.5 gives the result of accuracy by ensemble of methods with different seed values with dataset 1(DS1).

Table 5.2 DS1:Accuracy results by ensemble of methods having seed value 42

S.No.	Methods	Accuracy
1.	Recursive partitioning and regression tree	37.5
2.	Random Forest	37.5
3.	Generalized linear machine	58.3333
4.	Recursive partitioning and regression tree + Random Forest	43.47826087
5.	Recursive partitioning and regression tree + Generalized linear machine	60.86956222
6.	Random forest + Generalized linear machine	60.8695652
7.	Recursive partitioning and regression tree + Random Forest + Generalized linear machine	65.2173913

Table 5.3 DS1:Accuracy results by ensemble of methods having seed value 30

S.No.	Methods	Accuracy
1.	Recursive partitioning and regression tree	37.5
2.	Random Forest	41.66667
3.	Generalized linear machine	66.66667
4.	Recursive partitioning and regression tree + Random Forest	50
5.	Recursive partitioning and regression tree + Generalized linear machine	75
6.	Random forest + Generalized linear machine	70.83333
7.	Recursive partitioning and regression tree + Random Forest + Generalized linear machine	75

Table 5.4 DS1:Accuracy results by ensemble of methods having seed value 52

S.No.	Methods	Accuracy
1.	Recursive partitioning and regression tree	33.333333
2.	Random Forest	50
3.	Generalized linear machine	62.5

4.	Recursive partitioning and regression tree + Random Forest	54.166666667
5.	Recursive partitioning and regression tree + Generalized linear machine	66.66666667
6.	Random forest + Generalized linear machine	66.6666667
7.	Recursive partitioning and regression tree + Random Forest + Generalized linear machine	70.83333333

Table 5.5 DS1:Accuracy results by ensemble of methods having seed value 60

S.No.	Methods	Accuracy
1.	Recursive partitioning and regression tree	45.83333
2.	Random Forest	45.83333
3.	Generalized linear machine	45.83333
4.	Recursive partitioning and regression tree + Random Forest	54.16667
5.	Recursive partitioning and regression tree + Generalized linear machine	66.6666667
6.	Random forest + Generalized linear machine	62.5
7.	Recursive partitioning and regression tree + Random Forest + Generalized linear machine	66.666666667

The comparative analysis of accuracy of prediction for dataset 1 with cross validation of results with different seed values is shown in figure 5.1-5.4

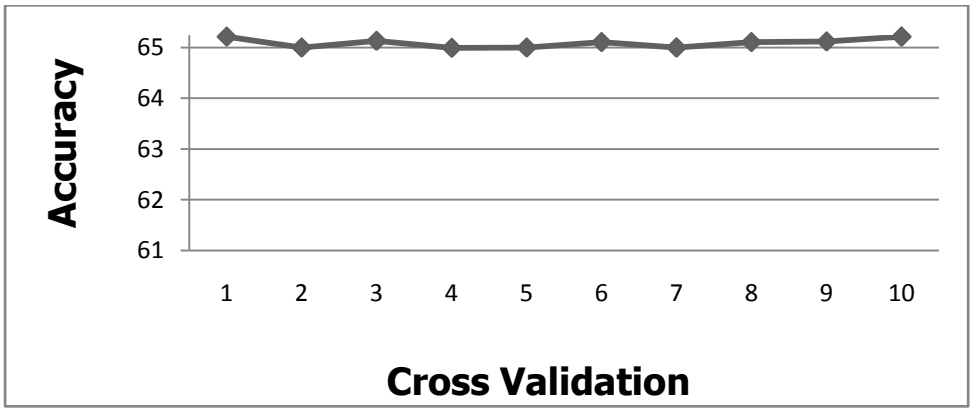


Figure 5.1 DS1:Comparative analysis of accuracy with cross validation with seed value 42

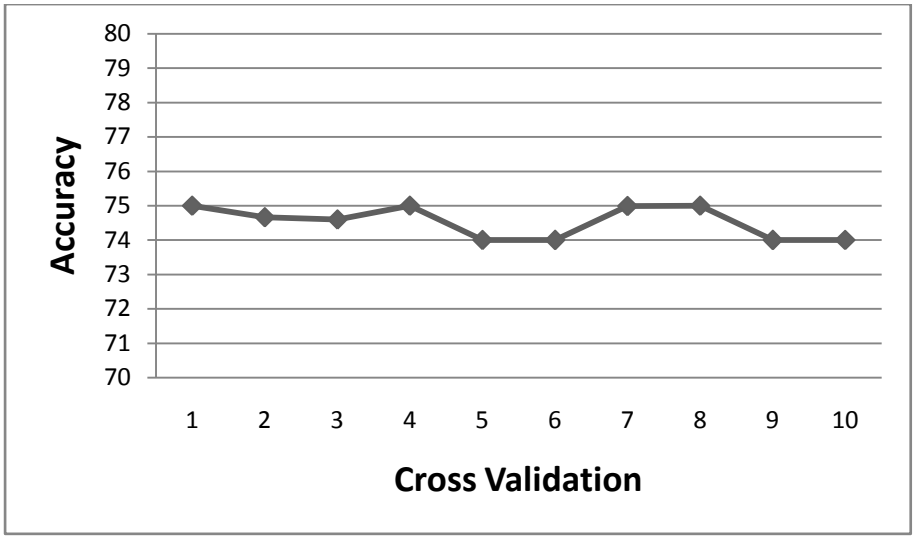


Figure 5.2 DS1:Comparative analysis of accuracy with cross validation with seed value 30

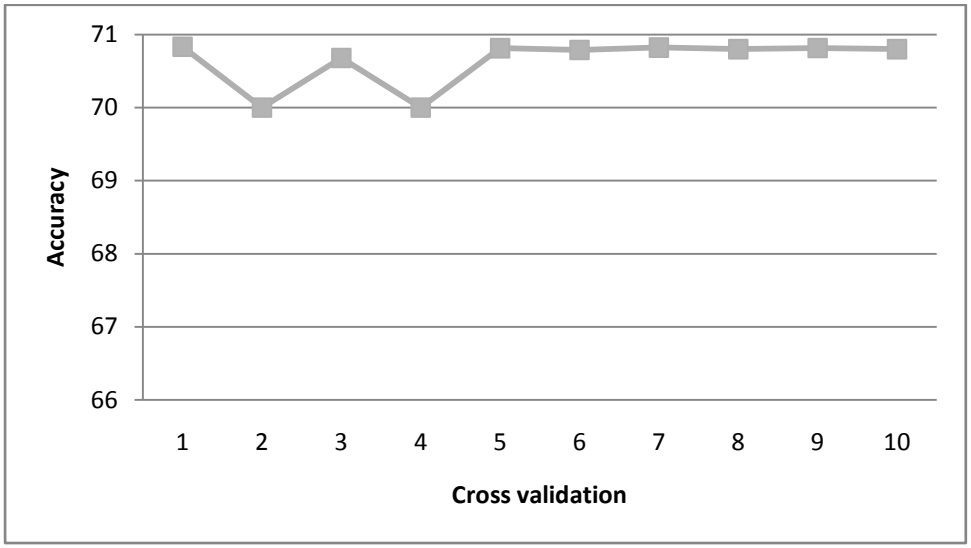


Figure 5.3 DS1:Comparative analysis of accuracy with cross validation with seed value 52

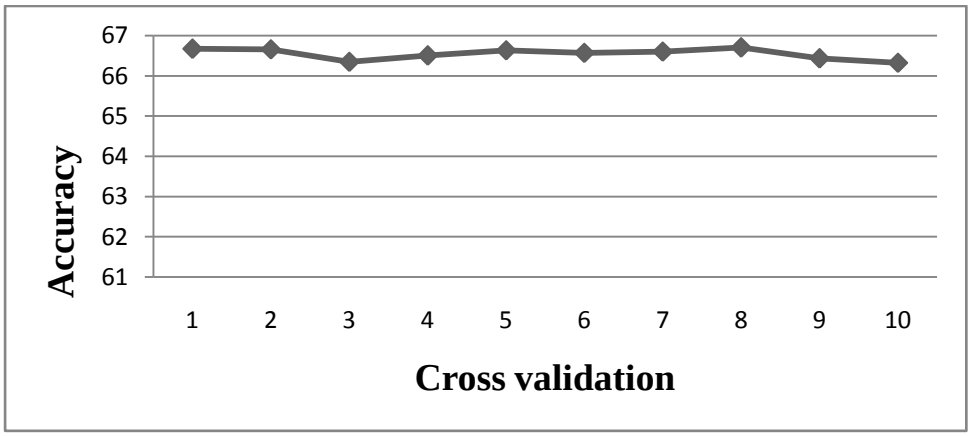


Figure 5.4 DS1: Comparative analysis of accuracy with cross validation with seed value 60

Similar results of second dataset (DS2) are shown in tables 5.6-5.9

Table 5.6 DS2:Accuracy results by ensemble of methods having seed value 42

S.No.	Methods	Accuracy
1.	Recursive Partitioning and Regression Tree	63.3333
2.	Random Forest	73.33333
3.	Kernlab Support Vector Machine	70
4.	Generalized Linear Machine	76.66667
5.	Recursive Partitioning and Regression Tree + Random Forest	80
6.	Recursive Partitioning and Regression Tree + Kernlab Support Vector Machine	76.66667
7.	Recursive Partitioning and Regression Tree + Generalized Linear Machine	86.6667
8.	Random Forest + Kernlab Support Vector Machine	76.66667
9.	Random Forest + Generalized Linear Machine	86.66667
10.	Kernlab Support Vector Machine + Generalized Linear Machine	83.33333
11.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine	80
12.	Recursive Partitioning and Regression Tree + Random Forest + Generalized Linear Machine	90
13.	Recursive Partitioning and Regression Tree + Generalized Linear Machine + Kernlab Support Vector Machine	86.6667
14.	Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	86.66667
15.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	90

Table 5.7 DS2:Accuracy results by ensemble of methods having seed value 30

S.No.	Methods	Accuracy
1.	Recursive Partitioning and Regression Tree	73.3333
2.	Random Forest	73.3333
3.	Kernlab Support Vector Machine	70
4.	Generalized Linear Machine	76.6667
5.	Recursive Partitioning and Regression Tree + Random Forest	80
6.	Recursive Partitioning and Regression Tree + Kernlab Support Vector Machine	83.3333
7.	Recursive Partitioning and Regression Tree + Generalized Linear Machine	90
8.	Random Forest + Kernlab Support Vector Machine	80
9.	Random Forest + Generalized Linear Machine	83.333
10.	Kernlab Support Vector Machine + Generalized Linear Machine	80
11.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine	83.3333
12.	Recursive Partitioning and Regression Tree + Random Forest + Generalized Linear Machine	86.66667
13.	Recursive Partitioning and Regression Tree + Generalized Linear Machine + Kernlab Support Vector Machine	90
14.	Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	86.66667
15.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	90

Table 5.8 DS2:Accuracy results by ensemble of methods having seed value 52

S.No.	Methods	Accuracy
1.	Recursive Partitioning and Regression Tree	56.6667
2.	Random Forest	66.6667
3.	Kernlab Support Vector Machine	70
4.	Generalized Linear Machine	70
5.	Recursive Partitioning and Regression Tree + Random Forest	70
6.	Recursive Partitioning and Regression Tree + Kernlab Support Vector Machine	73.333
7.	Recursive Partitioning and Regression Tree + Generalized Linear Machine	80
8.	Random Forest + Kernlab Support Vector Machine	76.6667
9.	Random Forest + Generalized Linear Machine	76.6667
10.	Kernlab Support Vector Machine + Generalized Linear Machine	83.333
11.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine	76.6667
12.	Recursive Partitioning and Regression Tree + Random Forest + Generalized Linear Machine	83.333
13.	Recursive Partitioning and Regression Tree + Generalized Linear Machine + Kernlab Support Vector Machine	80
14.	Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	83.3333
15.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	83.33333

Table 5.9 DS2:Accuracy results by ensemble of methods having seed value 60

S.No.	Methods	Accuracy
1.	Recursive Partitioning and Regression Tree	83.3333
2.	Random Forest	70
3.	Kernlab Support Vector Machine	63.3333
4.	Generalized Linear Machine	63.33
5.	Recursive Partitioning and Regression Tree + Random Forest	93.333
6.	Recursive Partitioning and Regression Tree + Kernlab Support Vector Machine	93.333
7.	Recursive Partitioning and Regression Tree + Generalized Linear Machine	93.3333
8.	Random Forest + Kernlab Support Vector Machine	73.333
9.	Random Forest + Generalized Linear Machine	76.6667
10.	Kernlab Support Vector Machine + Generalized Linear Machine	73.333
11.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine	73.333
12.	Recursive Partitioning and Regression Tree + Random Forest + Generalized Linear Machine	93.3333
13.	Recursive Partitioning and Regression Tree + Generalized Linear Machine + Kernlab Support Vector Machine	93.333
14.	Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	80
15.	Recursive Partitioning and Regression Tree + Random Forest + Kernlab Support Vector Machine + Generalized Linear Machine	93.3333

The comparative analysis of accuracy of prediction for second dataset with cross validation with different seed values is shown in figure 5.5-5.8

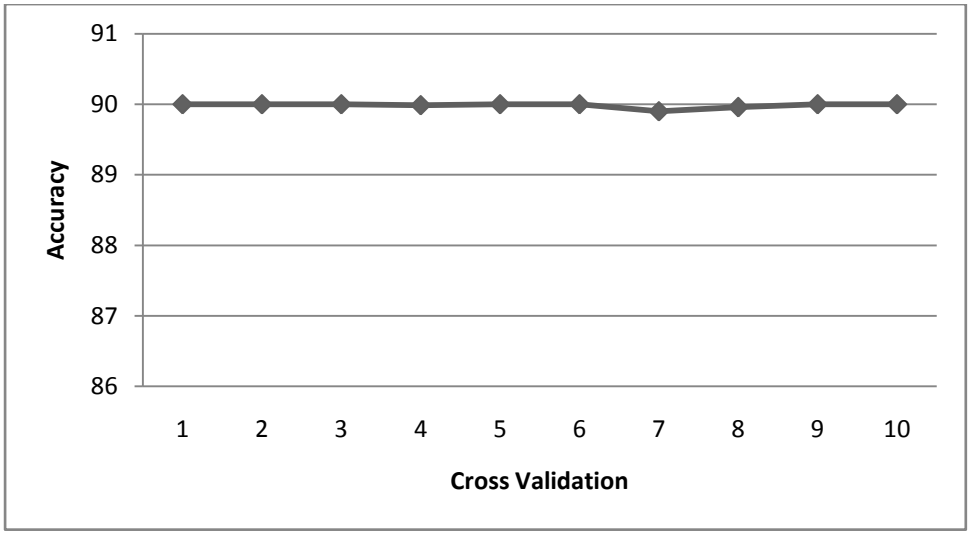


Figure 5.5 DS2:Comparative analysis of accuracy with cross validation with seed value 42

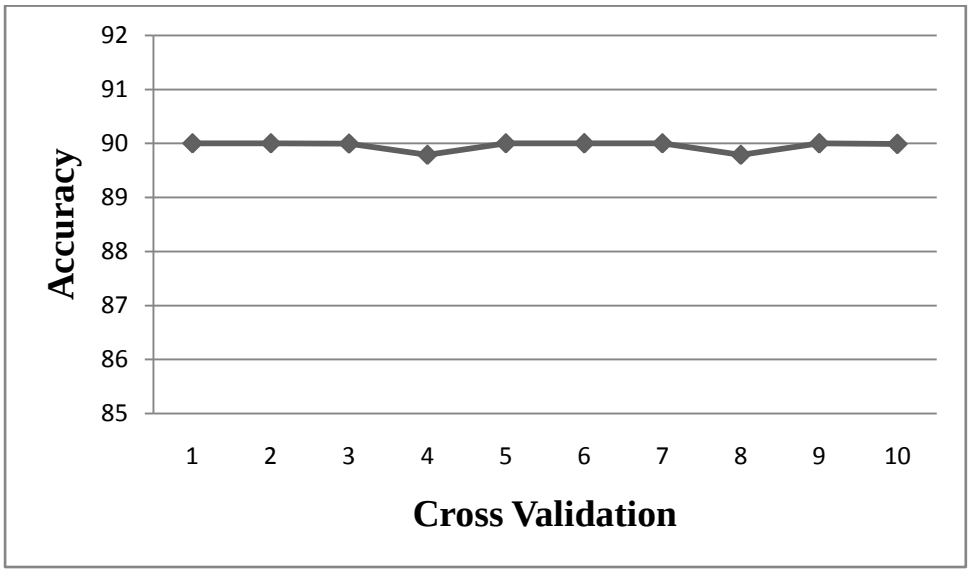


Figure 5.6 DS2:Comparative analysis of accuracy with cross validation with seed value 30

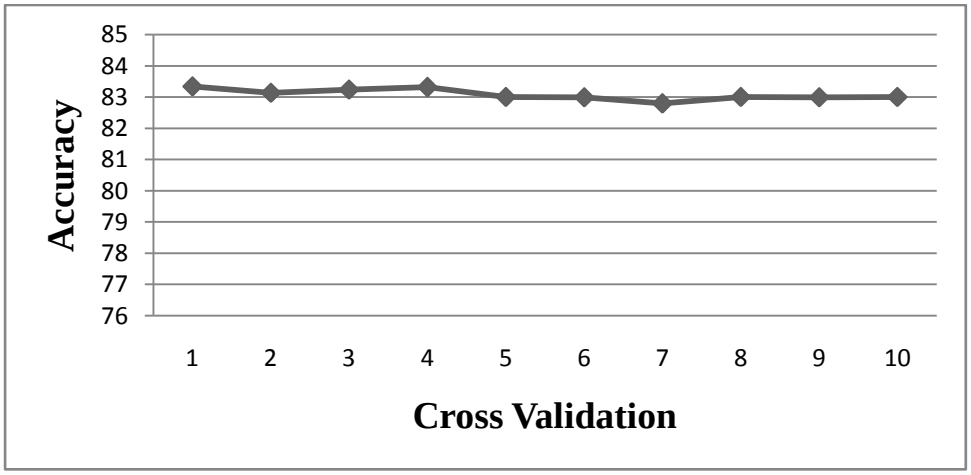


Figure 5.7 DS2:Comparative analysis of accuracy with cross validation with seed value 52

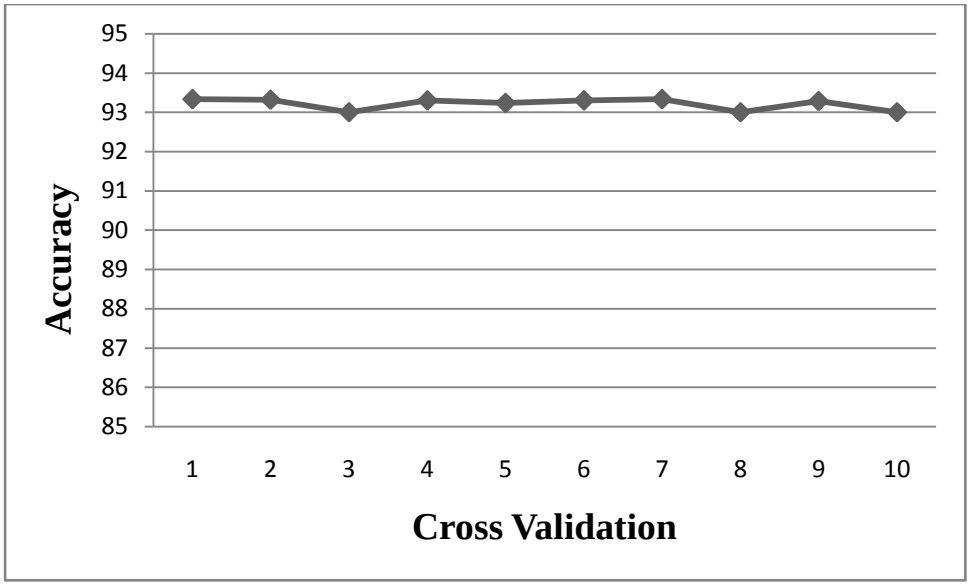


Figure 5.8 DS2:Comparative analysis of accuracy with cross validation with seed value 60

Discussion:

In previous part, there are results of experiments carried out on two datasets to get desired results. As we have combination of various models with different performance. These experiments are done with k fold validation to check the consistency of the system. The k value taken is 10 and the seed values are 42, 30, 52 and 60.

For first dataset, in table 5.2 the most efficient results are given by rpart, glm and rf together having seed value 42. In table 5.3, seed value 30 we get the best results by rpart and glm. In table 5.4 having seed value 52, the more efficient results are given by all the three algorithms collectively. In table 5.5 having seed value 60 the best results are given by rpart and glm. But if we want to conclude in the manner that the results are taken by only one seed value then the system will respond well with seed value 30 and the methods used for results are rpart and glm.

For second dataset, in table 5.6 having seed value 42 the best results are given by rpart, rf and glm. In table 5.7 the best results are given by rpart, glm and ksvm with seed value of 30. In table 5.8 the best results are given by glm and ksvm and seed value is 52. In table 5.9 the results are shown for the seed value of 60 which gives best results with highest accuracy of all. And most of the ensemble methods are giving the good results.

CHAPTER 6 CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Over the last decade, the ensemble based systems have enjoyed a growing attention and popularity in many applications due to their properties. There is a wide range of areas where there is a need of classifying the data items. There is diversity in the ensemble modelling as we have a large number of available methods which can be used for experiment. There are a large number of combination rules which can be applied in different areas of research. There is a huge scope of using these factors in various fields.

The work in this thesis comprises of methods used in applications so to improve the accuracy of the system by combining two or more models. The first dataset is a company dataset which comprises of class label which defines their sector and this sector can be predicted by system. System is trained by the training data and then it learns from it. Next the system will respond to the data which is going to test for future and the validation can be checked through this. The results are based on features in the dataset and the attribute values. The accuracy of system has been evaluated by the output responded by the system. Different techniques responded differently and have variable accuracy. So to have efficient results there is option to consider two or more models at a time. Similarly in second dataset there is class label of the career option to be taken by the individual. To improve the accuracy to classify and predict, the proposed model uses ensemble classifier and has proved to be successful for different seed values.

6.2 Future Scope

There is no single ensemble generation algorithm or combination rule that is universally better than others. All of the approaches discussed above have been shown to be effective on a wide range of real world and benchmark datasets, provided that the classifiers can be made as diverse as possible. The work done here can be extended upto many applications in any field. There may be the fields such as medical and finance where there is great need of the accuracy as compared to time, so the ensemble modelling concept comes as it is directly related to welfare and security of individual. As an example, in medical field it is necessary to give best result to predict

the disease as if it is related to someone life so we need consistent and accurate results. It is required to accurately predict the chances of bank crash etc.

REFERENCES

[1]. A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, vol. 31, pp. 264-323, 1999.

[2]. X. Rui_, D. Wunsch, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, vol. 16, pp. 645 – 678, 2005.

[3]. J. Kleinberg, "An impossibility theorem for clustering", *Conf. Advances in Neural Information Processing Systems*, vol. 15, pp. 463–470, 2002.

[4]. O. A. Abbas, "Comparisons between Data Clustering Algorithms", *International Journal of Information Technology*, vol. 5, pp. 320-325, 2008.

[5]. S. B. Kotsiantis, P. B. Pintelas, "Recent Advances in Clustering: A Brief Survey", *WSEAS Transactions on Information Science and Applications*, vol. 1, no. 1, pp. 73–81, 2004.

[6]. Xu Rui, C. W. Donald II, "Survey of Clustering Algorithms", *IEEE Transactions on Neural Networks*, vol. 16, pp. 645-678, 2005.

[7]. A. K. Jain, "Data Clustering: 50 Years Beyond K-Means", *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010.

[8]. S.Vijiyarani, S. Sudha, "Disease Prediction in Data Mining Technique-A Survey", *International Journal of Computer Applications and Information Technology*, vol. 2, no. 1, pp. 17-21, January 2013.

[9]. E. Osmanbegovic, M. Suljic, "Data Mining Approach for Predicting Student Performance", *Economic Review – Journal of Economics and Business*, vol. 10, no. 1, pp. 3-12, May 2012.

[10]. M.S.B. PhridviRaj , C.V. GuruRao, "Data mining – past, present and future – a typical survey on data streams", *Procedia Technology*, vol. 12 , pp. 255 – 263, 2014.

[11]. S. S. Sansgiry, M. Bhosle, K. Sail, "Factors That Affect Academic Performance Among Pharmacy Students", *American Journal of Pharmaceutical Education*, vol. 70, no. 5, Article 104, 2006.

[12]. Q. Radaideh, E. Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", *IJACSA*, vol. 3, pp. 144-151, 2012.

[13]. H. K. Kriegel, K. M. Borgwardt, P. Kröger, A. Pryakhin, M. Schubert, A. Zimek (2007) "Future trends in data mining*", Data Mining and Knowledge Discovery,* vol. 15, pp. 87– 97, 2007.

[14]. T. Velmurugan, "Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data", *Applied Soft Computing*, vol. 19, pp. 134–146, 2014.

[15]. E. W. T. Ngai, Hu Yong, Y. H. Wong, Y. Chen, X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", *Decision Support Systems,* vol. 50, pp. 559-569, 2011.

[16]. Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", *Acsys CRC, CSIRO,* 1998.

[17]. L.V. C. André, F. Everlândio, K. Faceli, "Multi-objective design of hierarchical consensus functions for clustering ensembles via genetic programming", *Decision Support Systems*, vol. 51, pp. 794-809, 2011.

[18]. B. Aviad, G. Roy, "A decision support method, based on bounded rationality concepts,to reveal feature saliency in clustering problems", *Decision Support Systems*, vol. 54, pp. 292–303, 2012.

[19]. Sandeep, Priyanka, R. Bansal, "Performance Comparison of Various Partition based Clustering Algorithms", *IJEMR,* pp. 216-223, 2014.

[20]. C. Combes, J. Azema, "Clustering using principal component analysis applied to Autonomy–disability of elderly people", *Decision Support Systems*, vol. 55, pp. 578–586, 2013.

[21]. Q. Zhao, P. Fränti, "WB-index: A sum-of-squares based index for cluster validity", *Data & Knowledge Engineering*, vol. 92, pp. 77–89, 2014.

[22]. O. J. Oyelade, O. O. Oladipupo, I. C. Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", *IJCSIS*, vol. 7, pp. 292-295, 2010.

[23]. G. N. Rao, M. Ramachandra, "A Study on the Academic Performance of the Students by Applying K-Means Algorithm", *IJETCAS*, pp. 14-180, 2014.

[24]. Y. Liu, Li Qianqian, X. Tang, Ma Ning, R. Tian, "Superedge prediction:What opinions will be mined based on an opinion supernetwork model", *Decision Support Systems*, vol. 64, pp. 118–129, 2014.

[25]. A. Adhikari and P. R. Rao, “Efficient clustering of databases induced by local patterns”, *Decision Support Systems*, vol. 44, pp. 925–943, 2008.

[26]. A. Irpino, R. Verde, Francisco de A.T, Carvalho, “Dynamic clustering of histogram data based on adaptive squared Wasserstein distances”, *Expert Systems with Applications*, vol. 41, pp. 3351–3366, 2014.

[27]. P. L. Lin, P. W. Po-Huang, P. H. Kuo, Y. H. Lai, “A size-insensitive integrity-based fuzzy c-means method for data clustering”, *Pattern Recognition*, vol. 47, pp. 2042–2056, 2014.

[28]. F. U. Xiao, C. Fan, “Data mining in building automation system for improving building operational performance”, *Energy and Buildings*, vol. 75, pp. 109–118, 2014.

[29]. J. Jacques, C. Preda, “Model-based clustering for multivariate functional data”, *Computational Statistics and Data Analysis*, vol. 71, pp. 92–106, 2014.

[30]. L. D. Angelis, J. G. Dias, “Mining categorical sequences from data using a hybrid clustering method”, *European Journal of Operational Research*, vol. 234, pp. 720–730, 2014.

[31]. J. Demšar, B. Zupan, “Orange: Data Mining Fruitful and Fun - A Historical Perspective”, *Informatica*, vol. 37, pp. 55–60, 2013.

[32]. S. Srivastava, “Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining”, *International Journal of Computer Applications* (0975 – 8887), pp. 88-110, 2014.

[33]. C. Romero,S. Ventura, “Educational data mining: A survey”, *Expert Systems with Applications*, vol. 33, pp. 135–146, 2007.

[34]. J. S. Breese, D. Heckerman, C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, *Microsoft Research*, Morgan Kaufmann Publishers, pp. 1-18, 1998.

[35]. *R. I. D. Baker, K. Yacef, “*The State of Educational Data Mining:A Review and Future Visions*”, JEDM - Journal of Educational Data Mining*, vol. 1, pp. 3-16, 2009.

[36]. S. Padmaja and S. S. Fatima, “Opinion Mining and Sentiment Analysis – An Assessment of Peoples’ Belief: A Survey”, *International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC)*, vol. 4, pp. 21-33 no. 1, 2013.

[37]. R. Basili, Di Nanni, M. and M. T. Pazienza, "Engineering of IE systems: an object oriented approach", *In: Pazienza, editor, Information Exctraction,* LNAI 1714, pp. 134–164, 1999.

[38]. D. Ferrucci, A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research Environment", *Natural Language Engineering*, vol. 10, pp. 327 – 348, 2004.

[39]. Y. Low, J. Gonzalez, A. Kyrola, A. Bickson, C. Guestrin, U. C. Berkeley(2010) "GraphLab: A New Framework For Parallel Machine Learning" , 2010.

[40]. N.S.V. Rao, "On fusers that perform better than best sensor", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 8, pp. 904–909, 2001.

[41]. J.R. Quinlan, "Bagging, boosting and C4.5, *13th Int. Conf. on Artificial Intelligence*, pp. 725–730, 1996.

[42]. G. Fumera and F. Roli, "A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 942–956, 2005.

[43]. Various Authors, "Proceedings of International Workshop on Multiple Classifier Systems (2000–2005)", F. Roli, J. Kittler, T. Windeatt, N. C. Oza, and R. Polikar, Eds. Berlin, Germany: Springer, 2005.

[44]. T.K. Ho, "Random subspace method for constructing decision forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[45]. B.V. Dasarathy and B.V. Sheela, "Composite classifier system design: Concepts and methodology", *Proceedings of the IEEE*, vol. 67, no. 5, pp. 708–713, 1979.

[46]. D.H. Wolpert, "Stacked generalization", *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[47]. L.I. Kuncheva, "Change-glasses' approach in pattern recognition", *Pattern Recognition Letters*, vol. 14, no. 8, pp. 619–623, 1993.

[48]. K. Woods, W.P.J. Kegelmeyer, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 4, pp. 405–410, 1997.

[49]. R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton, "Adaptive mixtures of local experts", *Neural Computation*, vol. 3, pp. 79–87, 1991.

[50]. M.J. Jordan and R.A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation*, vol. 6, no. 2, pp. 181–214, 1994.

[51]. L. Xu, A. Krzyzak, and C.Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[52]. T.K. Ho, J.J. Hull, and S.N. Srihari, "Decision combination in multiple classifier systems", *IEEE Trans. on Pattern Analy. Machine Intel.*, vol. 16, no. 1, pp. 66–75, 1994.

[53]. G. Rogova, "Combining the results of several neural network classifiers", *Neural Networks*, vol. 7, no. 5, pp. 777–781, 1994.

[54]. L. Lam and C.Y. Suen, "Optimal combinations of pattern classifiers", *Pattern Recognition Letters*, vol. 16, no. 9, pp. 945–954, 1995.

[55]. I. Bloch, "Information combination operators for data fusion: A comparative review with classification", *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 26, no. 1, pp. 52–67, 1996.

[56]. S.B. Cho and J.H. Kim, "Combining multiple neural networks by fuzzy integral for robust classification", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no.2, pp. 380–384, 1995.

[57]. L.I. Kuncheva, J.C. Bezdek, and R. Duin, "Decision templates for multiple classifier fusion: An experimental comparison", *Pattern Recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[58]. F. Smieja, "Pandemonium system of reflective agents", *IEEE Transactions on Neural Networks*, vol. 7, no. 1, pp. 97–106, 1996.

[59]. H. Drucker, C. Cortes, L.D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and other ensemble methods", *Neural Computation*, vol. 6, no. 6, pp. 1289–1301, 1994.

[60]. L.I. Kuncheva, "Classifier ensembles for changing environments", 5th Int. Workshop on Multiple Classifier Systems, *Lecture Notes in Computer Science*, F. Roli, J. Kittler, and T.Windeatt, Eds., vol. 3077, pp. 1–15, 2004.

[61]. L.I. Kuncheva, *Combining Pattern Classifiers, Methods and Algorithms*. New York, NY: Wiley Interscience, 2005.

[62]. E. Alpaydin and M.I. Jordan, "Local linear perceptrons for classification", *IEEE Transactions on Neural Networks*, vol. 7, no. 3, pp. 788–792, 1996.

[63]. F. Roli, G. Giacinto, and G. Vernazza, "Methods for designing multiple classifier systems", 2nd Int. Workshop on Multiple Classifier Systems, in *Lecture Notes in Computer Science*, J. Kittler and F. Roli, Eds., vol. 2096, pp. 78–87, 2001.

[64]. G. Giacinto and F. Roli, "Approach to the automatic design of multiple classifier systems", *Pattern Recognition Letters*, vol. 22, no. 1, pp. 25–33, 2001.

[65]. L. Breiman, "Bagging predictors", *Machine Learning*, vol. 24, no. 2, pp. 123–140,1996.

[66]. R.E. Schapire, "The strength of weak learnability", *Machine Learning*, vol. 5, no. 2, pp.197–227, 1990.

[67]. Y. Freund and R.E. Schapire, "Decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[68]. F.M. Alkoot and J. Kittler, "Experimental evaluation of expert fusion strategies," *Pattern Recognition Letters*, vol. 20, no. 11–13, pp. 1361–1369, Nov. 1999.

[69]. J. Kittler, M. Hatef, R.P.W. Duin, and J. Mates, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[70]. K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition*, vol. 29, no. 2, pp. 341–348, 1996.

[71]. G. Fumera and F. Roli, "A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 942–956, 2005.

[72]. L.I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, 2002.

LIST OF PUBLICATIONS

1. Mansi Gera and Shivani Goel, "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity", in *International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015, pp. 22-29*
2. Mansi Gera and Shivani Goel, "A Model for Predicting the Eligibility for Placement of Students Using Data Mining Technique", in proceedings of IEEE Xplore digital Library on "International Conference on Computing, Communication and Automation" (ICCCA-2015), Galgotias university, Greater Noida, May-2015
3. Mansi Gera and Shivani Goel, "Predicting the Profession from User Background using Ensemble Modeling" in ICCIT-2015 (Communicated)

VIDEO PRESENATATION

The video link is available on

https://www.youtube.com/watch?v=VBz_F47ZTnE&feature=em-share_video_user

REFLECTIVE DIARY

January

In order to study about the domain for my research, to have deep knowledge about it, I started studying the vast topic in December 2014 and in the month of January, I was able to start preparing with a survey paper. This study constituted the literature review done by me on main domain of my topic i.e. Data Mining.

Study of papers which are published by various authors in SCI, IEEE and others journals were read by me to have clear view of the topic and through this I was aiming to get clarity with my problem statement which could be identified to be solved.

February

As we know research never ends. It goes on increasing day by day. Anyone can go into more and deep level of search about any particular topic of his or her interest. After surveying I came up with the survey on tools of data mining which are used to implement various techniques of data mining. A research paper on the survey of data mining tools, techniques, methods and algorithm details was started for communicating to a journal. Tools were compared on a large number of features and areas of application.

By doing survey I was somehow clear with the pros and cons of the algorithm. There was a clear vision of the applicability of the techniques in various fields, i.e. how could we use these techniques in specific domain and where there are to be applied. Side by side I was preparing with the summary of the research problems which were solved by the researchers in this field.

After this I came to know about one of most interesting phase of data mining - Predictive analysis. Predictive analysis is done for predicting the future events. So I went for the application related to this sub-domain. To carry out my work with predictive domain, I went on to deeper research and this resulted in understanding various steps required to complete this.

March

In the month of March, I was aiming to submit a research paper in IEEE conference. As we came with the field prediction analysis and the technique to solve this i.e.

classification, we continued with the study and solution procedure and were able to implement our first model. This model was the prediction for students. For our initial experiment and implementation, we took the data of our own university on small scale (sample size 100) so that this work will be carried in future to again come up with some advance concept. The dataset of placement records was collected by our own university and the reason behind this was to improve the scenario of placements of students. Data collection from placement cell, analyzing this data and pre processing for cleansing the data was done.

To implement our first step we used classification model. Then the results were recorded and tested with students of our university. With the implementation part my guide gave me directions to write this part into one paper and we communicated it to IEEE International Conference on Computing, Communication and Automation (ICCCA-2015), at Galgotias University, Greater Noida, to be held in May-2015.

I also got my first research paper on "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity" accepted and published in International Journal of Computer Applications (0975 – 8887) in vol.113, no. 18, with pages 22-29.

April

After the implementation of model and carrying out further study, we again came up with new concept in which we can combine various methods into one method. And the approach used to solve this problem is mixture of experts i.e. ensemble of classifiers. As we know prediction is all about the accuracy as if we are predicting any future event, then it is predicted accurately. So I discussed this idea with my guide and she appreciated this idea and gave me more details about it and I started studying.

After this I came to know my latest research area where I actually decided to work. And this area is called as Ensemble of classifiers models. There are more than 50 models which are used for classification of data. And for ensemble of these models we have many hybrid rules which are used to combine these. So the through study about ensemble modelling was carried out by me to enrich my knowledge regarding this. I started learning R language.

So we were able to decide the research problem on which we were going to work. The language in which it was to be implemented was decided. The notification of

acceptance of my second research paper communicated to ICCCA 2015 was received in April 2015.

May

In this time of the semester, the implementation part of the research work was carried out. In the mid of month I attended the international conference ICCCA 2015 at Galgotias University, Greater Noida in Up, India and presented my paper. I also noted various comments and suggestions for my future work. It was great to listen to many researchers at the conference and discussing their areas of work.

June

I completed the implementation part of the proposed algorithm and complied all the results. The work was communicated as a research paper in IEEE ICCIT-2015.

I completed all required details in my thesis report. All hard work done was finally in the form of my thesis report. I thank God and my parents for providing me strength and capability to do all what I could do for my Master of Engineering. I am also thankful to my guide for her support and all my fellow colleagues for helping me out at odd hours.

(Mansi Gera)

(25^{th} June, 2015)