1. Introduction

The field of Human Resource Management has been very strongly influenced by today's highly competitive and globalized economy. New trends emphasize human resource as a prime resource for the success of an organization. Thus there has been a need to develop new models for effective management of intellectual and human capital which in turn may expedite the decisions for intelligent investment in the human resource. HRM basically is a set of tasks to maintain and develop competent human resource and is based on comprehensive analysis of employee data. In this study, data mining techniques are applied to employee data with an objective to discover interesting patterns and relationships. This chapter is dedicated to HRM overview, statement of the problem and the objective of the study and concludes with data mining – its basic concepts and application and a brief introduction to supervised and unsupervised learning techniques in DM.

1.1 HRM – An Overview

Over the years, there has been realization by the organizations that improvement in technology and infrastructure alone cannot improve the overall performance of the organization. Organization's people are its knowledge power. This knowledge power is the most important driving factor for success. Thus, as a result of drastically changing business environment, human resource and consequently human resource management have taken center stage.

1.1.1 HRM Functions

Human Resource Management (HRM) is defined as the art of procuring, developing and maintaining competent workforce to achieve the goals of an organization in an effective and efficient manner. HRM is not a one shot activity but is a process. The objective of the process is to bring people and organizations together for achieving the goals of each.

Even though people have always been central to organizations, in today's knowledge based industries, their strategic importance is growing radically. KSAs – knowledge, skills and abilities of employees has become a decisive factor in the success of any organization. KSAs help establish a set of core competencies that distinguish an organization from others. HRM has also undergone evolution in recent years. It now functions based on new line of thinking that a team

of competent and committed employees delivers the goals if they are involved in important activities and are encouraged to develop the goals that they are supposed to achieve. Thus, HRM has a vital role to play in formulating the strategic plan as well as in the execution of strategies.

HRM functions can be broadly categorized as managerial and operative. Basic managerial functions include

- Planning
- Organizing
- Directing
- Controlling

Operative functions include

- Procurement Functions
- Development
- Motivation and compensation
- Maintenance
- Integration
- Emerging Issues

Fig 1.1 shows the diagrammatic representation of typical activities carried out by HRM

Technological advancement and globalization in recent past has made it essential for the organizations to be more competitive and fast. HRM plays a crucial role for achieving competitive advantage through people by adopting appropriate HR policies, strategies and practices. This is done by focusing on quality, customer service, employee involvement teamwork and productivity.

In the above mentioned process, a large amount of data is generated and needs to be maintained. Human resource Information System (HRIS) is a data system that caters to the informational needs of HRM. It is used to collect, analyze and report information about people and job.

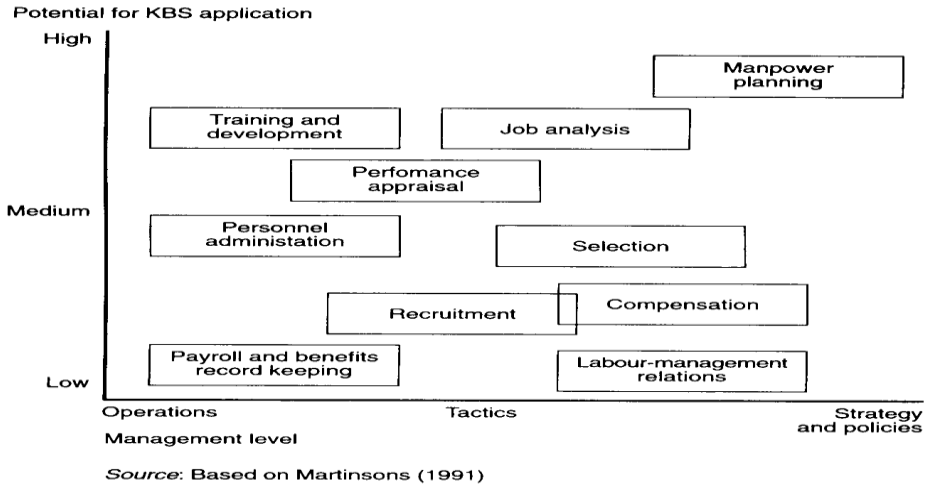


Figure 1.1 Functions of HRM

1.1.2 Performance Appraisal Process

A performance appraisal, employee appraisal, performance review, or (career) development discussion is a method by which the job performance of an employee is evaluated (generally in terms of quality, quantity, cost, and time) typically by the corresponding manager or supervisor.

It is a method of evaluating an individual's work performance in order to arrive at objective personnel decisions. Appraisal is carried out periodically and according to a definite plan. Employee appraisal results are useful in

- Compensation decisions
- Promotion decisions
- Training and development programmes
- Feedback
- Personal development

Thus, employee performance evaluation forms a basis in many HRM decisions. Even though the concept of formal appraisal process is very young, informally the practice of appraisal is considered to be an ancient art. But, in the absence of structured appraisal system, ethical, legal and motivational issues may arise.

Employee performance evaluation, one of the most important activities of HRM, is systematically carried out using following sequence of steps

1. Establish Performance Standards
2. Communicate the standards
3. Measure and actual performance
4. Compare actual performance with standards and discuss the appraisal
5. Taking corrective actions if necessary

Thus, for an effective and bias free performance evaluation process, transparency in the process is of utmost importance and steps 2, 3 and 4 are implemented accordingly. Step 1 is about setting up standards related to a particular job.

Performance Appraisal forms

For every job various tasks are carried out by an employee, and performance refers to the degree of accomplishment of these by the employee. The performance evaluation process is an analysis of various abilities and weaknesses of the employee that are observed while executing the job. Every organization has predefined criteria for assessing its employees based upon the job. Performance appraisal forms are the tools for performance evaluation. A typical form contains a list of KPIs and a rating scale to rate the employee. For Example Knowledge, execution, Productivity etc are some of the factors included in the appraisal form as KPI. The list of such factors is usually varies from job to job and organization to organization. Different types of rating scales are also possible, like checklists, weighted checklists, graphical rating scales etc. With the help of the responses given on the rating scale, an employee's performance is measured. The ratings could also be explained for better evaluation. For Example a typical 5 point numeric scale can be elaborated further as -

EXCEPTIONAL (5): Consistently exceeds all relevant performance standards. Provides leadership, fosters teamwork, is highly productive, innovative, responsive and generates top quality work. Active in industry-related professional and/or community groups.

EXCEEDS EXPECTATIONS (4): Consistently meets and often exceeds all relevant performance standards. Shows initiative and versatility, works collaboratively, has strong technical & interpersonal skills or has achieved significant improvement in these areas.

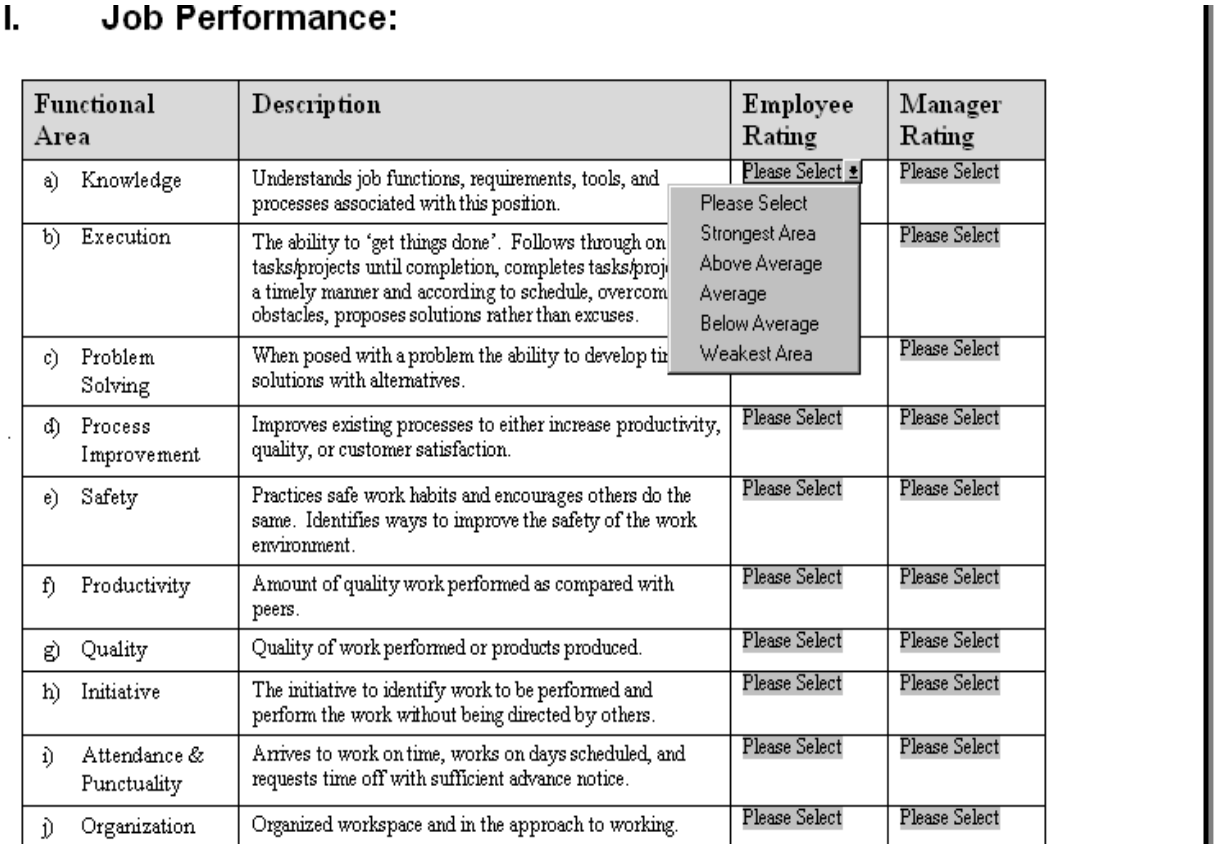
MEETS EXPECTATIONS (3): Meets all relevant performance standards. Seldom exceeds or falls short of desired results or objectives. Lacks appropriate level of skills or is

inexperienced/still learning the scope of the job.

BELOW EXPECTATIONS (2): Sometimes meets the performance standards. Seldom exceeds and often falls short of desired results. Performance has declined significantly, or employee has not sustained adequate improvement, as required since the last performance review or performance improvement plan.

NEEDS IMPROVEMENT (1): Consistently falls short of performance standards.

Fig 1.2 shows a typical appraisal form



I. Job Performance:

Functional Area	Description	Employee Rating	Manager Rating
a) Knowledge	Understands job functions, requirements, tools, and processes associated with this position.	Please Select	Please Select
b) Execution	The ability to 'get things done'. Follows through on tasks/projects until completion, completes tasks/proj a timely manner and according to schedule, overcom obstacles, proposes solutions rather than excuses.		Please Select
c) Problem Solving	When posed with a problem the ability to develop ti solutions with alternatives.		Please Select
d) Process Improvement	Improves existing processes to either increase productivity, quality, or customer satisfaction.	Please Select	Please Select
e) Safety	Practices safe work habits and encourages others do the same. Identifies ways to improve the safety of the work environment.	Please Select	Please Select
f) Productivity	Amount of quality work performed as compared with peers.	Please Select	Please Select
g) Quality	Quality of work performed or products produced.	Please Select	Please Select
h) Initiative	The initiative to identify work to be performed and perform the work without being directed by others.	Please Select	Please Select
i) Attendance & Punctuality	Arrives to work on time, works on days scheduled, and requests time off with sufficient advance notice.	Please Select	Please Select
j) Organization	Organized workspace and in the approach to working.	Please Select	Please Select

Please Select
Strongest Area
Above Average
Average
Below Average
Weakest Area

Figure 1.2 Sample Appraisal Form

In the above shown appraisal form, skills are represented under the heading area and two columns for rating are provided. Thus the filled form will contain responses for a set of skills by appraise (The employee) and the appraiser (Supervisor).

1.2 Objective of the Study

Every bit of data generated by any process in an organization may represent important information. KDD is all about discovering such hidden knowledge and make the concerned process more effective by facilitating decision making. Data mining techniques thus provide competitive edge by automatically extracting useful knowledge from the data. Because of it's the interdisciplinary nature DM has always been a field with potential research scope. In recent years researchers and analysts have been exploring the suitability of DM techniques in different domains by proposing novel DM models.

Formal Performance evaluation or appraisal of an employee is generally done annually, mostly for the purpose of salary review. Performance of employees can have impact on many important decisions like deciding appropriate training programs, allocation of right person to right job or planning career paths for employees. The performance evaluation process is an analysis of various abilities and weaknesses of the employee. Every organization has predefined criteria for assessing its employees. For example knowledge, execution, productivity, interpersonal skills etc. HRM of the organization carries out performance evaluation in a defined and structured way.

The objective of this study is to analyse the past employee performance evaluation data (appraisal records) using DM techniques to –

- Extract inter relationships among various parameters that contribute the performance of an employee.
- Predict the performance category of an employee using the extracted information.
- Group the employees based on inherent similarities among them.
- Enhance the prediction process by integrating supervised and unsupervised learning methods.

The study aims at developing a classifier model which could act as a decision support tool for HR personnel while taking decisions based on performance of employees.

1.3 Problem Statement

Organizational competitiveness in today's economy is directly proportional to the quality of its human capital or intellectual capital. Thus an employee in any organization is the focus point of all HRM activities in order to excel in terms of cost, quality and innovation. The effective

management of the human resource is the need of the day. HRM basically is a set of tasks to maintain and develop competent human resource. There has been a need to develop new models for effective management of intellectual and human capital which in turn may facilitate the decisions for intelligent investment in the human resource.

The performance evaluation process is an analysis of various abilities and weaknesses of the employee. Many important HRM decisions like deciding appropriate training programs, allocation of right person to right job or planning career paths for employees are based on performance assessment of employees

Every organization has predefined criteria for assessing its employees. For Example Knowledge, execution, Productivity etc can be some of the factors included in the evaluation process. The list of such factors is usually very extensive. For each of the factor in the evaluation criteria, the employee is rated on some given scale. Thus the performance evaluation is the extensive analysis of large number of attribute values which generally is carried out annually. Till now the main purpose of this activity used to be salary review. In recent past the along with the evolution in HRM, the perspective of evaluation process is also changing. Along with being related to salary reviews, performance evaluation is now seen as a main activity which may significantly contribute in maintaining and developing a competent human resource. Availability of an "intelligent" decision support can be beneficial in this scenario. Such tool would essentially perform analysis of past employee data to reveal interesting relationships among the data. Thus a decision maker would be provided with right information at right time, thereby enhancing HR activities.

The study entitled "Employee Performance Prediction Model" uses data mining techniques for analysis of above mentioned data with an objective of predicting employee performance. The model developed during the course of the study justifies the suitability and effectiveness of proposed methods as compared to traditional tools and techniques currently used in the process.

1.4 Data Mining in Knowledge Discovery Process

In the recent past, the capabilities of generating and collecting data have been increasing significantly. Computerization of business, scientific and government transactions, use of digital cameras etc, are the driving factors for this explosive growth of the data. This abundant data is a repository of hidden and interesting information. Data mining which is also called as knowledge discovery is the process of extracting interesting patterns and correlations from large amount of

data. Though data mining primarily is intelligent analysis of data, it is an umbrella term which encompasses different activities based on the different contexts and domains.

Data mining is a multidisciplinary and evolving field based on the advents in database technology, statistics, pattern recognition, information retrieval, machine learning, high performance computing and data visualization and several other disciplines. Various data mining techniques aim at the discovery of patterns hidden in large data sets focusing on issues related to their feasibility, effectiveness, usefulness and scalability by generating suitable models.

1.4.1 Knowledge Discovery Process

Databases collect and store tremendous amount of data generated as a result of day to day transactions. Thus, without sophisticated analysis and retrieval tools, the operational databases are described as data rich but information poor repositories Knowledge discovery or knowledge mining can thought of as an automated process of finding hidden, novel and useful patterns that facilitates the transformation of a database into a knowledge base. Knowledge discovery is an iterative set of specific tasks and activities carried out in a predefined sequence. A typical knowledge discovery process begins with data preparation and ends with the discovery of and presentation of extracted knowledge. Even though the terms knowledge discovery and data mining are used synonymously, data mining is essentially is a step in the knowledge discovery process. One or more data mining techniques can be used for extracting ‘information’ or ‘knowledge’ from the operational databases. Knowledge discovery as a process consists of following steps –

- Problem statement defines the business objective of the knowledge discovery process and determines the expected operational environment of the result.
- Data selection retrieves the task relevant data.
- Data cleansing removes any noisy data, handles missing values, and removes irrelevant and redundant attributes and tuples.
- Data integration combines the data from multiple, possibly heterogeneous data sources and makes the data consistent and unified.
- A data transformation performs necessary transformations like discretizing attribute values, changing the layout of the data tables, construction of new (derived) attributes.
- Data mining applies an appropriate algorithm to discover patterns.

- Pattern evaluation visualizes discovered patterns, lets domain experts examine the patterns and identify interesting, useful, and valid patterns; Also, identifying and discarding trivial, questionable, or invalid patterns is performed.
- Pattern utilization uses the discovered patterns to achieve the business objective defined at the beginning of the process.

Diagramatic representation of a typical knowledge discovery preocess id depicted in the fig 1.1

Data mining is at the heart of the above mentioned KDD process. With the help of its rich and efficient techniques and algorithms it extracts information and outputs interesting patterns.

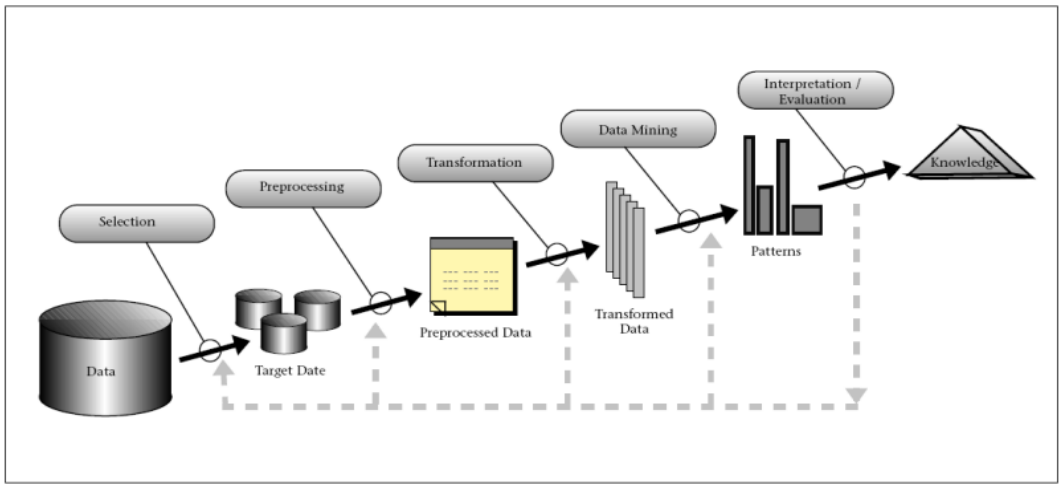


Figure 1.3 Steps in a KDD process

Different categories of data mining tasks performed to analyze the data and "mine" information are given below -

- Association rule discovery is finding sets of elements frequently appearing together in the database. The frequent items sets are then used to generate association rules based on the statistical tests of their co-occurrences. With the availability of several variations to the basic process, mining frequent items and association rules can be advantageous in deciding correlated item sets and interesting relationships among complex data.
- Sequential patterns mining is finding sequences of events which preserve a given order in many sequences stored in the database. Each sequence is used to generate a rule describing the most common ordering of events.

- Classification and prediction is one of the most commonly used DM techniques. It is the processes of building a model which can predict unknown values of class attribute (known as class label). It is a two-step process. In the first step a classifier is constructed based on tuples with class labels (Training the classifier). Second step uses, this classifier model for classification of tuples with unknown class attribute values. This process is also categorized under supervised learning techniques because the model is based upon data with previously known class values. The process of classification is described further in detail in section
- Clustering is process of finding similarities of data objects and dividing objects into disjoint groups such that the intra-group similarity of objects is maximized and, at the same time, the inter-group similarity is minimized. In contrast to the classification clustering does not take any input for grouping the data and hence is called as unsupervised learning.
- Characterization is extraction of general features of the dataset and creating descriptions that can be used to summarize properties of the original data.
- Discrimination is the extraction of features which differentiate a target group of data objects from a contrasting class to better illustrate the features of a target group as opposed to the contrasting class.
- Outlier detection is the process of identifying the data objects which are very different from the expectations that they should be treated in a special way. Such data objects might result from errors, noise, or they could represent a typical or exceptional cases

1.4.2 Data Mining – A Knowledge Acquisition Tool

The availability of historic data makes data mining as one of the most effective analysis techniques in many domains. Typical data mining application domains are -

- Financial Data Analysis – Typical operations which could be supported using DM for financial analysis include loan payment prediction, classification and clustering of customers for marketing and detection of financial crimes.

- Retail Industry – Suitability of DM techniques for the operations like multidimensional analysis of sales, analysis of effectiveness of campaigns, customer retention analysis has been well established.
- Telecommunication industry – Telecommunication industry could be benefitted by using DM for several analysis tasks like fraudulent pattern analysis, multidimensional and sequential pattern analysis
- Biological Data Analysis – The tremendously growing field of genomics, proteomics an allied areas of biomedical research are posing challenges and requirements for suitable analysis techniques for the complex structures that they work on. In the recent years "Bioinformatics" thus uses a set of dedicated DM techniques called as biological data mining which are used with complex structures like protein and DNA sequences. Common DM tasks include similarity search in protein sequences, integration of proteomic and genomic databases, discovery of structural patterns and analysis of genetic networks etc.
- Business Intelligence - In recent years, business intelligence has emerged as a major field which uses data mining techniques for analysis. Business intelligence (BI) is formally defined as "set of concepts and methodologies to improve decision making in business through the use of facts or fact based systems." The main objective of BI projects is to gain sustainable competitive advantage in business. It is an umbrella term used collectively for activities, tools and techniques used in the process of delivering "right information" to right people on right time. For transforming raw data to right information, DM techniques are used.

Other Areas of Applications

- Medicine - drug side effects, hospital cost analysis, genetic sequence analysis, prediction etc.
- Scientific discovery - superconductivity research, etc.
- Engineering - automotive diagnostic expert systems, fault detection etc

Since mining methods can add intelligence to the analysis, newer domains are also exploring the suitability of DM techniques thus can be used in every domain as a knowledge acquisition tool.

1.5 Thesis Organization

The thesis is a report of the study carried out during the process of the research work. It describes the various methods and techniques used in the study. The thesis is organized in five chapters to present the stepwise progression of the study which led to the findings.

Chapter 2 discusses background work done in the domain of data mining for HRM. It highlights different models developed and suggested by researchers for effective applications of DM in HRM. It briefly presents the methodologies used in these models. Since these models involve several sub processes, the chapter continues with a brief review of related techniques like data preprocessing, missing value handling and feature extraction. The chapter concludes with an overview of classification and clustering algorithms.

Chapter 3 is dedicated to the methodology followed in the study. Starting with the brief outline of the steps followed in building the classification model, the chapter continues with the detailed discussion of every step. It highlights the preprocessing techniques used in the study, and the different approaches followed to develop a supervised classifier for performance prediction. Application of unsupervised learning technique clustering and its significance is also described. The chapter concludes with a discussion of classification algorithms and tools (Weka and MATLAB) used for data analysis.

Chapter 4 summarizes results obtained in each step of the study. It discusses performance of the different classification models for parameters namely accuracy percentage, ROC statistics and confusion matrix. Based on the observations, the chapter further continues with result discussions and concludes with important observations of the study.

Chapter 5 is the last chapter of the thesis. It first summarizes the inferences concluded from the results and discusses limitations and challenges faced by the researcher during the course of the study. It concludes with a brief introduction of future work and enhancements that can be made.

2. Literature Review

The field of data mining is very much interdisciplinary in nature. Systems based on knowledge discovery techniques are very commonly used in many fields like healthcare, bioinformatics and market analysis. The applications of data mining techniques to human resource management are relatively new. HRM, even today uses statistical analysis tools for most of the analysis. Because of the evolutions in the HRM, there is a continuous need for newer model for more effective functioning of HRM. Consequently, data mining researchers are experimenting with this less explored domain of human resource management. This chapter discusses basic concepts, current practices, tools and techniques used for employee performance evaluation and prediction. It also describes the work done by various researchers in the field of Data mining and HRM.

2.1 Background Work

HRM collects and generates huge amount of data while performing its activities. With sophisticated database management techniques in place, the maintenance and retrieval of this data has become a non-issue in the recent past. This data is with several complexities and nonlinearities among variables. To analyze this data, common statistical techniques like discriminant analysis and regression are less efficient as compared to the data mining tools such as neural networks, genetic algorithms, decision trees, fuzzy logic, and rough sets. The abundance of data has attracted data mining research towards the domain of Human Resource Management. Data mining techniques are aimed at discovering knowledge from the available data and could be used for improving the processes. Stephen Baker, in his article “Data Mining Moves to Human Resource”, (Bloomberg BusinessWeek, Predictive Analytics, March 12, 2009) has provided some insights to strengthen above initiatives. He writes, “As the role of computers in the workplace expands, employees leave digital trails detailing their behaviour, their schedule, their interests, and expertise. For executives to calculate the return on investment of each worker, their human resources departments are starting to open their doors to the quants.” The article also highlights the inclusion of data mining approaches in the statistical software like SAS for the analysis of employee data. The author has mentioned about Microsoft studying correlations between successful workers and the schools and companies they arrived from. Also analysis of communications within Microsoft was also performed. These types of analysis provide new insights to the organizations (especially in this case HRM). Large corporations have begun coupling analytics to evaluate their workforces. The objective of individual analysis may vary but

the ultimate goal is to improve the productivity of an organization by understanding hidden patterns in employee performance.

Employee performance evaluation is very crucial to several strategic decisions. It also forms the basis for deciding training needs of an employee, planning career paths for an employee, allocating right people for right job etc. Thus performance evaluation is central to many of HR activities. A tool which could analyze employee data to discover useful patterns from the data might facilitate the decision making process. The extraction of hidden information could be done by applying one or more data mining techniques and may provide important insights into interrelationships between various parameters that contribute to performance of an employee. Such knowledge could significantly contribute to decision making by making it intelligent and more effective. The rest of this section is an overview of the work done by researchers in the above mentioned context.

Performance evaluation and personnel selection are the most widely studied topic from HRM domain by data mining researchers. The importance of performance evaluation and the process has been explained in earlier sections. Both these activities involve an analysis of various skills possessed by an employee for the execution of a job. Organizations collect and maintain this data in a predefined format .Every skill or KPI (Key performance Indicator) represent a data attribute. Thus, based on the level of proficiency shown by the employee for each of the skills in the format, the employee is rated on a scale. According to traditional approach, Final rating could be a simple mathematical formula which calculates a value indicating the score of the employee [1]. But there could be interrelationships present between different data attributes. For example all the employees having rating good for innovation also have rating good for interpersonal skills. Such and more complex relationships and patterns can be extracted from the data.

2.1.1 A Generic DM Model for HRM

There have been models suggesting the effectiveness of data mining techniques in HRM [2]. These models are built using mainly classification and prediction, clustering and association mining. Some of the objectives of these models include -

- Predict the percentage accuracy in employee performance
- Predicting employee's behavior and attitude
- Analyze forecast and model information to quantify human capital assets

- Predicting the performance progress throughout the performance period
- Identifying the best profile for different employees

Each of the work cited uses a generic knowledge discovery model [8] essentially having five steps -Data understanding, Data preparation and identification of DM technique, Data mining, Evaluation and Knowledge consolidation. The consolidated knowledge then would be applied to the problem domain.

2.1.2 Talent Forecasting Model for HRM

Even though classification and prediction are commonly used data mining techniques for employee performance evaluation, applying unsupervised techniques like clustering is also suggested. [3] Have suggested a model for talent management that can be used as a decision support tool and performs several different type of analysis. The model uses supervised as well as unsupervised techniques. The authors have further tested the accuracy of prediction using four different classification algorithms and have achieved significant accuracy in the classification results. The approach and the steps are described in the following diagram.

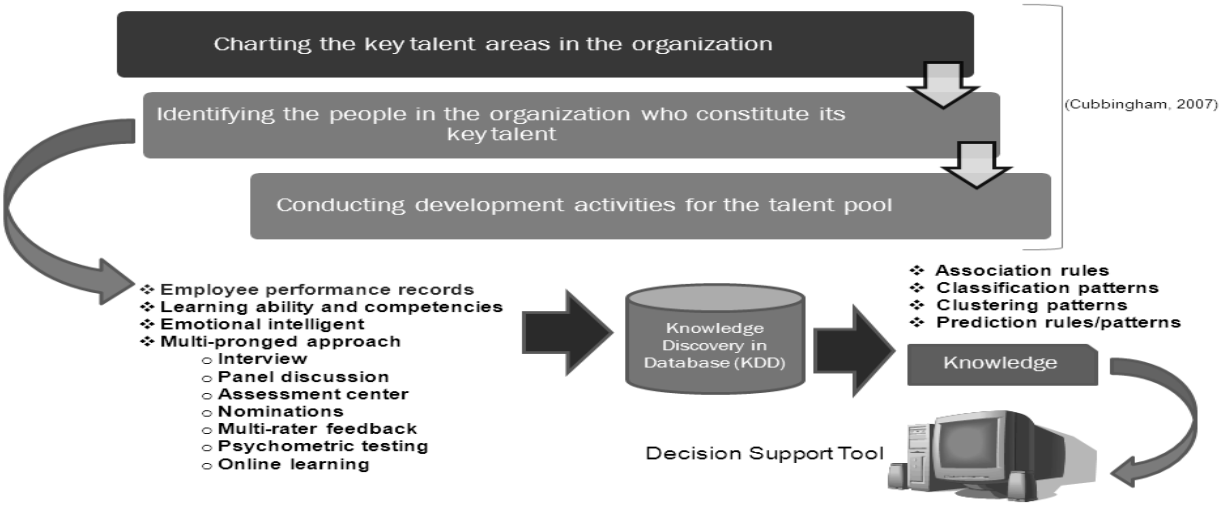


Figure 2.1 DM for talent management [3]

2.1.3 Employee Performance Classification – Data warehousing approach

[1] has also built similar model but with data warehousing perspective of the data. Additionally, the author has used clustering as one of the preprocessing step for feature selection. The methodology followed by this model can be represented as a sequence of steps -

1. Data Selection – Selection of task relevant data
2. Data Preprocessing – For unification of data values, ranges and data type consistency
3. Cluster Analysis – To understand the interrelationships between the attributes for better classification accuracy
4. Attribute Subset Selection – To identify and retain attributes with high significance
5. Classification – Using decision tree, formulating the rules for classification

2.1.4 Employee Clustering – Fuzzy Approach

A very similar task of assessing and selecting right people can be simplified by applying appropriate data mining techniques.[4] But, the above described models do not take into account any uncertainty or vagueness that may inherently exist in the available information. Fuzzy data mining and rough set theory are the data mining techniques that can handle vagueness and uncertainty of data by incorporating fuzzy set theory into mining algorithms. Fuzzy set theory allows us to deal with vagueness by offering high levels of abstraction. Fuzzy data mining algorithms are used in various application domains. Rough set approach is more suitable for noisy and imprecise data. It is an extension of mathematics concept of equivalence classes with an inclusion of approximation.

[5] has proposed fuzzy data mining approach for employee selection where fuzzy clustering algorithm has been applied to group employees. Fuzzy algorithms are based on the concept of degree of belongingness of an object. Fuzzy clustering makes use of fuzzy distance calculation and similarity matrix. In contrast to the classical clustering, in fuzzy clustering, an element may belong to more than one cluster with a different degree of belongingness. This approach of grouping provides additional insights into the skillset and characteristics of the employee particularly when the data available is uncertain and imprecise.

2.1.5 Work Behavior Prediction – Rough Set Approach

[4] Has proposed dynamic work and job analysis for talent management. Authors have claimed that predicting work behavior including resignations can be effectively performed with the application of rough set approach to allow approximation in the process.

2.2 Review of Related Techniques and Methods

As mentioned in section 1.4 knowledge discovery expected in these above described models is not a one-step output but is a sequence of steps. This section provides a brief overview of literature referred during different stages of the study.

2.2.1 Data Pre processing

Data preprocessing is one of the initial steps in generic KDD process (Section 1.3) is very important since "dirty" and un prepared data may have negative impact on the results of a DM process. Imputation, which is a process of filling missing values, is the very first step in data preprocessing. There has been work going on in this area and several new techniques and variations to the basic approaches have been proposed. Unsupervised techniques like simple competitive learning [8] provide advantages over simple mean based statistical imputation. Along with generic imputation techniques, methods for specific algorithms have been proposed. [9], is such supervised imputation approach, which is based on fuzzy learning methods and finds missing values for neural network algorithms.

Data generalization or discretization is also one of the extensively studied fields in data mining. Discretization is a very important preprocessing for applying classification techniques to a dataset containing numeric values with large range. Primitive discretization approaches like equidistance and equi frequency might hamper the quality of a classification algorithm. Supervised techniques proposing the inclusion of correlation between the generated intervals could be effective for discretization.[10]. The DBCHIMERGE algorithm proposed in [10] is also capable of elimination of irrelevant attributes simultaneously. [11] Also uses the correlation structure present in the dataset for discretization to develop unsupervised discretization model in

multivariate context. Rough set based discretization techniques are proposed [12] to handle the inherent approximation efficiently.

2.2.2 Dimensionality Reduction

Dimensionality reduction or feature extraction is a process of generating a reduced representation of the original dataset. Attribute subset selection is one such dimensionality reduction technique. It is a process of removing redundant and irrelevant attributes. As depicted in figure 2.2, it is the process by which a minimum set of attributes is derived such that the resulting probability distribution of the classes is as close as possible to the original distribution obtained using all attributes.

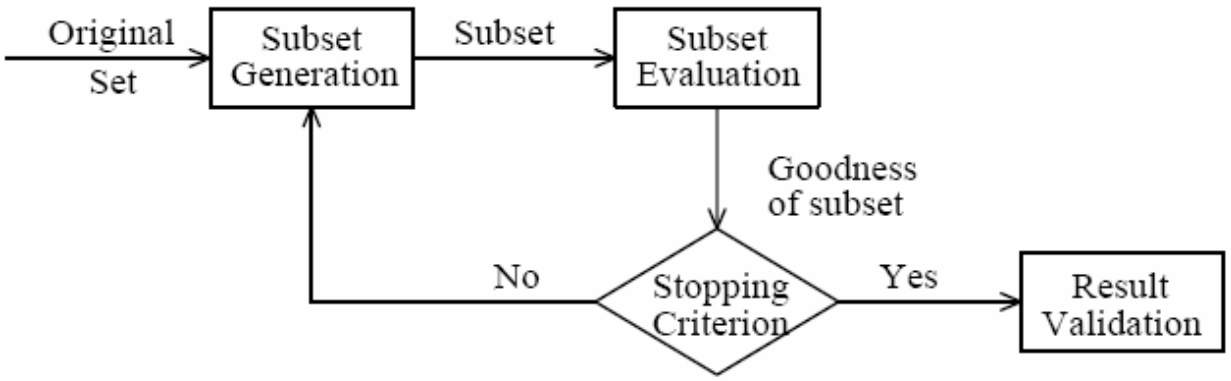


Fig 2.2 Attribute Subset Selection [13]

Attribute selection is a process in which a subset of M attributes out of N is chosen, complying with the constraint M≤ N, in such a way that characteristic space is reduced according to some criterion [13].There are two main reasons to keep the dimensionality of the features as small as possible: cost minimization and classification accuracy. Cost minimization is achieved because after feature selection the classifiers will be computed faster and will use less memory [12].

As with all other sub processes of KDD, attribute subset selection also is widely studied. While [14] gives an insight into early days correlation based feature selection methods. [16] is an overview of several modern feature selection approaches. Presently, there are two approaches for building attribute subset wrapper and filter.[13] Wrapper model approach uses the method of classification itself to measure the importance of features set, hence the feature selected depends

on the classifier model used. [16]. Different from the wrapper approach, the filter approach attempts to build an attribute subset independent of the classification algorithm being used. It considers only the data set for evaluation of the subset.

While wrapper methods generate high quality attribute subset, their performance for high dimensionality data is considerably low. Additionally generated subset is classifier specific and might not give same accuracy if used for different classifiers[13]. Filter approach produces a generic subset that could be used with any classifier. The evaluation of intermediate subsets is done based on measures like consistency and relevance. Since it follows a generic evaluation approach accuracy of prediction may vary from classifier to classifier[13]. Also there is a possibility of correlated attributes [15]. To combine the advantages of the two approaches hybrid models are developed. Clustering is extensively used and a natural choice for feature selection. Basic clustering techniques can be combined with other techniques to get better results[15,17]. Models also been also proposed which extends the unsupervised learning method to supervised using the rough set approximation approach. [1] Has also used clustering as primary feature selection method in his model of performance prediction.

2.3 Review of Supervised and Unsupervised Learning Methods

Each of the work cited in the section 2.1 primarily has used classification - supervised learning approach or clustering – unsupervised learning for constructing DM models in HRM. Next two sections present basic concepts and a review of popularly used algorithms for above-mentioned techniques. Advantages and limitations and of every technique haves also been highlighted.

2.3.1 Supervised Learning Technique – Classification

"Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances." [18]. Typical steps of a supervised learning approach are shown in the figure 2.3. It aims at building a concise model of the distribution of class labels in terms of predictor features. Every step in the process is very significant and affects the quality of performance of the results obtained. If the performance of the classifier built is not satisfactory, then backtracking through the entire process might be required. Previous sections have highlighted the research work reviewed during every sub process of the study.

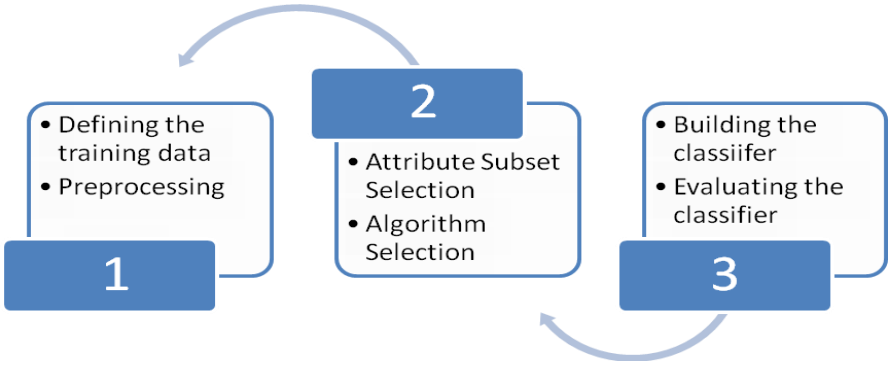


Figure 2.3 Steps of a supervised learning method based on [18]

Classification is a data analysis task of constructing a classifier to predict categorical labels called as (class labels). For example the objective of the study is to predict the performance category of an employee such as A, B or C. Thus the class labels are discrete values. A predictor on the other hand predicts a numerical value and the task of building such model is called as prediction. Data classification is a two-step process- 1) Learning step which is also called as training phase. In this phase a classifier is built based on previously known data set of data classes (training set). Thus the classification algorithm learns from the training data set that essentially a set of data tuples with class labels. Since the class labels of the tuples in training set are made available to the algorithm, this step is called as supervised learning.

Data mining classification algorithm set includes different algorithms based on different types of data, different data characteristics and different problem domains. In the initial days of data mining, most of the classification algorithms encountered performance issues because of their characteristics of being memory resident. But, there have been enhancements in subsequent years to make the algorithms scalable. While there are advantages and limitations of every algorithm, classification and prediction continues to be the most dominating DM technique because of its rich set of algorithms.

Classification is an extensively used supervised learning technique DM technique. [18] gives a detailed review of classification algorithms and classifies the algorithms in five categories namely – logic based, perceptron based, statistical learning algorithms, case based learning and support vector machines.

Prominent in the category of logic based algorithms are decision tree and rule bases classifiers. A decision tree algorithm classifies instances based on values of features and a tree is constructed in the process [19]. The approach of tree construction by splitting dataset based on values of feature makes it suitable for univariate data[18]. Decision trees because of its effectiveness in handling category data, is still widely used in spite of other issues. Non-requirement of domain knowledge is also a very significant advantage of decision trees. Decision tree algorithms tend to be sensitive to the training set [19].

"A decision tree, or any learned hypothesis *h*, is said to overfit training data if another hypothesis *h′* exists that has a larger error than *h* when tested on the training data, but a smaller error than *h* when tested on the entire dataset" [18]. Overfitting of a decision tree is a common problem and approaches for minimizing overfitting of decision tree are suggested by researchers [19].

Rule based classifiers also are very easy to comprehend. Final set of rules is obtained after evaluation using statistical rule quality measure.

Neural network algorithms are perceptron based and are used widely because of various advantages like high tolerance to noisy data and the ability to classify patterns on which they are not trained. Well suited for continuous input and output, neural network algorithms are popularly used in the domain of [19]. The main disadvantage of neural networks is long training time hence the feasibility of using should be evaluated before using [18].

Statistical learning algorithms are based on the underlying probability structures. Bayesian networks are the most well known representatives of statistical learning algorithms. These algorithms have the advantages of significantly less training times and also well suited for category as well as numeric attributes. There have been attempts to overcome concerns related to conditional independence and several variations to the preliminary Bayesian model. The accuracy of Bayesian networks is still comparable with other logic-based algorithms like decision trees [20].

Instance based classifiers defer actual induction to the classification phase in contrast to other algorithms. This results in less training time but huge memory requirements. KNN is one of the powerful lazy learning algorithms which has been used effectively in many domains but has the disadvantages of memory requirements and sensitivity to the value of K[19].

Other newer approaches for classification include support vector machines which like neural network algorithms perform well with continuous input but cannot handle discrete data[18].

One of the very recent approaches in supervised learning methods is a set of algorithms incorporating imprecise data handling. These algorithms are fuzzy logic based or rough set theory based [20].

2.3.2 Unsupervised Learning - Clustering

Unsupervised learning techniques do not require the any external input like target or class variable. The only input required is the data set [21]. Clustering is a type of unsupervised learning technique that is used to analyze data sets in order to find out the natural structure and unknown but valuable patterns. Clustering is one of the extensively studied and commonly used DM techniques. Having several applications in many domains, clustering is commonly used as a feature selection technique. Clustering is also used as compression technique [22]. Above-mentioned models [3,1] also make used of clustering for improving classification process. [5] Has used fuzzy set theory and fuzzy clustering algorithm for grouping employees.
Like classification, clustering algorithms set also provides a variety of clustering solutions. Several categories of clustering algorithms like partitioning, hierarchical and density based are used according to problem and domain characteristics.

The novel data mining approach for performance analysis of employees is gradually becoming popular. Major reasons of this success could be highlighted as usefulness in understanding the performance in the context of all the parameters of all other employees instead of treating it as an individual’s attribute and providing intelligent analysis at right time by understanding the complex inter relationships

The above-mentioned models use both supervised and unsupervised techniques independent of each other but there has not been any attempt to integrate the results of these approaches. Such integration might be helpful in uncovering additional patterns and relationships among the attributes and might enhance the prediction process. Unsupervised learning results could be analyzed to understand inherent similarities among employee performance, to understand the effect of bias in the evaluation process etc.

Data Mining, with its evolving approach continues to contribute significantly to the KDD process in every domain. With the ongoing research in every sub process of KDD, latest developments in other disciplines like database technologies are being captured to cater the need of the problem domain. Since data mining is at the intersection of mathematics, statistics, database technology and machine learning, research in any of these disciplines stimulates DM research also. It is a field with huge research potential because of the interdisciplinary and ever progressing nature of DM. Depending on the problem and the data set a variety of primitive to advance and hybrid techniques are available. Along with application of DM techniques to newer domains like HRM, improvements to previously studied problems by applying newer techniques has also being explored. The dominance of rough set theory in recent research for every step of KDD to incorporate approximation (which might be inherent in many domains**)** is also noticeable**.**

3. Methodology

This chapter discusses the methods and approaches used in this research work. .The KDD process in the study was aimed at predicting performance of an employee. Different approaches and techniques used in every step of the process are discussed in brief. Starting with data description the chapter discusses the methods used in each step of the KDD to build the employee performance classification model.

3.1 Performance Classification Model – Block Diagram

Based on the complex and multi-level structure of the data, the generic process of KDD was reformed for effective results. The sequence of steps followed by the study is as follows-

1. Data Pre processing
2. Attribute subset selection – Direct, hierarchical and integrated approach
3. Build the classifiers – Direct , hierarchical and integrated
4. Evaluate the classifiers
5. Perform cluster analysis

Different steps of the methodology followed are diagrammatically represented in Figure 3.1

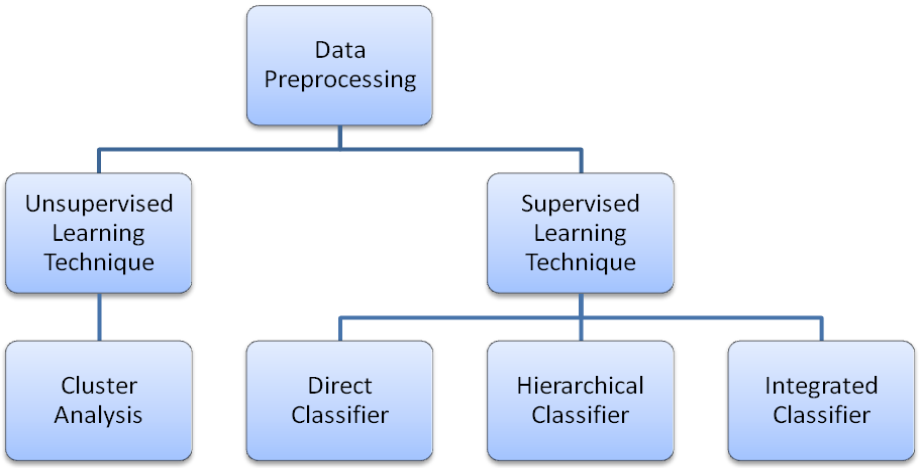


Figure 3.1 Steps of Building the Classifier(s)

3.2 Data Set Description

The study is carried out for data of an academic institution with faculty strength of over 300. The courses are divided under various deaneries. The institution evaluates the performance of a faculty member annually based on various predefined parameters that are included in an appraisal form. A very systematic appraisal process is in place, which is hierarchical in nature. Evaluation is performed based on responses given in appraisal forms by an employee and his/her supervisors. Thus the responses form the appraisal forms are mapped to a performance value using simple mathematical formulae. But the process does not take into account the complex relationships between the parameters contributing to the performance of an employee (A faculty member). The study was aimed at studying these interrelationships among the data parameters and predicting performance of an employee. It will further enhance the evaluation process by integrating supervised and unsupervised learning techniques to minimize inherent bias (if it exists) in the evaluation process.

The employee appraisal data (appraisal forms) collected has 133 different parameters (divided across employee and supervisors' evaluation forms) to be assessed for evaluating an employee. These parameters basically represent the skill set required to be possessed by a faculty member. In today's dynamic scenario, a teacher imparting higher education obviously has a greater role than merely taking classes. Thus, s/he is assessed based on several factors along with the proficiency in classroom teaching. The appraisal forms considered in the study include most of such attributes which would evaluate a faculty member appropriately according to his "ability" of execution of his role. The attribute ratings reflect this "ability" shown by the employee while executing the job. Along with the categorical attributes that are rated on three-point scale (A, B or C), other attributes like experience, research initiatives, and other achievements are also considered as appraisal data making the data very complex.

3.2.1 Data Characteristics

Since the data is of high dimensionality and is divides across different dataset, uniformity in naming the data attributes is essential. Following conventions are followed when refereeing to the attributes.

Attribute Name = Dataset name followed by attribute number within the dataset

The structure of appraisal data is which is divided in four data sub sets as follows-

- **Demographic Details of the employee – D1**

This data set contains details like name, age, experience, department etc. These attributes are numeric, nominal as well as descriptive.

Dataset D1 attributes are referred as
Attributes D1_1 to D1_14
Numeric Attributes 7
Nominal Attributes 4
Descriptive Attributes 3

- **Self-Evaluation Of the Employee – D2**

This data set contains categorical attributes that for which response is registered by the employee. This dataset represents self-evaluation of an employee. The attributes are further divided into subsections. Thus to assess particular ability, an elaborate list of attributes which are considered significant is also given making the no of attributes exhaustive.

Attributes D2_1 to D D2_59
Nominal Attributes 59

- **First appraiser evaluation – D3**

This data set is the rating responses of immediate supervisor of an employee. Most of the attributes are repeated in dataset 2 and dataset 3. Some additional attributes are also included which the appraiser uses to assess the employee in the capacity of the supervisor. Typically these attributes assess the faculty member for his/her capabilities in addition to class room teaching.
Attributes D3_1 to D3_50
Nominal Attributes – 50

- **Second appraiser evaluation – D4**

This dataset is essentially a consolidation of parameters from first three datasets. It does not contain individual parameter contained in the second and third dataset but contains only main skill parameter. Figure 3.1 shows typical parameters and how they are further detailed with sub

parameters making the structure of the appraisal form multi-level. Dataset D4 essentially contains parameters at the highest level. The sub parameters at the next level appear only in self-evaluation form and the first appraiser form.

D4_1 to D4_9 + Class Label Attribute

Nominal Attributes 9 + 1

The ratings given by second appraiser are translated to 5 point numeric scale and a score value is calculated which represents numeric performance value of the employee. This value is generalized in the study to represent performance category.

3.2.2 Multi-level Data Representation

A typical appraisal may be characteristically multi-level. The appraisal form shown in the figure 3.1 depicts a typical two-level data.

Attribute (Performance Factors)	**Possible Values**
Job Understanding 1. Understands job duties and responsibilities. 2. Possesses sufficient skill and knowledge to perform all parts of the job effectively, efficiently and safely. 3. Understands and promotes department mission and values. 4. Makes an active effort to stay current with new developments.	**Nominal** **(Rated on a scale of 1 – 5)** Categorical (Above average, average, Below Average)
Initiative/Innovation 1. Self-directed, resourceful, creative toward meeting job objectives. 2. Introduces new concepts and processes using independent and original thought	**Categorical** **(Above average, average, Below Average)** Categorical (Above average, average, Below Average)

Figure 3.2 Two-Level structure of appraisal data.

(Because of the confidentiality contract signed with organization, actual details of the data cannot be revealed).

Attributes are grouped under a main KPI attribute which further is elaborated using a set of sub meters. These sub parameters are also rated independently. Thus, the form above records one consolidated rating (shown in bold) and also ratings for (sub) parameters of the group level attribute.

The dataset used in the study also has similar features but evaluation of employee at three different levels add extra dimension to the analysis. This complex relationship among the datasets and the parameters is diagrammatically shown in the figure 3.3.

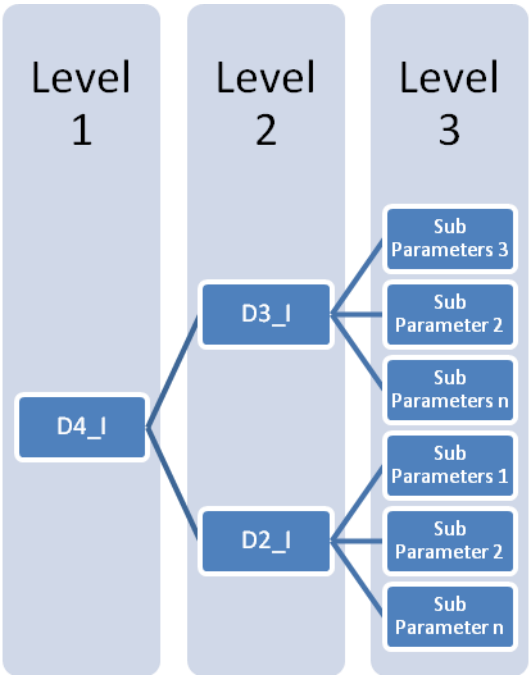


Figure 3.3 Hierarchical Representations of Performance Parameters

D4_I, a typical attribute of dataset 4 has corresponding group level parameters D2_I and D3_I as well as sub parameters in the datasets 2 and 3.

3.3 Data Understanding and Pre-processing

Data understanding is always the first step in any analysis. It is carried out to understand basic characteristics of the dataset. It also helps in selecting the techniques for subsequent steps in the analysis.

3.3.1 Correlation Analysis

Correlation analysis is a fundamental technique used to understand first level relationships between the parameters. It helps in identifying primary redundancies. Numeric correlation coefficient is calculated using the formula

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

Where,

Where x_i and y_i represent series of n n measurements of variables X and Y, $i = 1, 2, ..., n$,

x and y are the means of X and Y, s_x and s_y are the standard deviations of X and Y.

Correlation is a measure of the relation between two or more variables. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation.

Correlation analysis is particularly very significant in this study. Since an employee as well as supervisor(s) rate same parameter, to understand the relationship between the responses given by different hierarchy individuals, correlation coefficient was used.

To describe the relationship between numeric variables graphically scatter plots can also be used. A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern or trend between two numeric attributes. It is useful in providing an elementary understanding of outliers, clusters of points or possibility of correlation. A scatters plot treats each pair of values as a pair of coordinates in the plane and plots it accordingly. A scatter plot matrix is an extension of basic scatter plots and can be used to visualize each attribute with every other attribute.

3.3.2 Integration of datasets

Since the objective of the analysis is to find interesting relationships between the attributes, the datasets need to be combined. Since some attributes are repeated across the datasets correlation analysis was performed to identify and eliminate primary redundancies. All the attributes were retained because the correlation analysis performed did not reveal any redundancies. The repeated attributes were treated separately and were represented with separate names in the integrated dataset.

3.3.3 Filling Missing Values

After the basic understanding of the data, cleaning data is the next step performed. Since the real world data tend to be incomplete, filling these missing data is important. Several approaches for handling missing values like replacing with global constant, replacing with mean and replacing with class mean can be used. Following two approaches are used by the study to handle missing values based on their type

- Missing values in the dataset are handled by Weka's built in function which replaces missing a value by mean in case of a numeric value and by mode in case of a nominal value.
- Function knnimpute in MATLAB finds and replaces numeric missing value with its nearest neighbor value.

3.3.4 Data Generalization

It is the process of replacing raw data by higher-level concepts. Data generalization is also considered as one of the data reduction a technique that makes use of concept hierarchies. Data generation can be applied to numeric attributers (Example – age can be mapped to youth, middle-aged and senior) as well as categorical attribute (Example street could be mapped to higher level concept like city).

Data set D1 contains the demographic details of an employee. For effective classification process, these numeric attributes were generalized. Following attributes were generalized using MS Excel (Ranges are based on researcher's assumptions).

- Age

- Experience
- Research Experience
- Industry Experience
- Publication details
- Experience in the organization
- Performance value discretized as performance category

3.3.5 Representation of descriptive attributes as category/binary attributes

Dataset D1 also includes descriptive data about other achievements and responsibilities of an employee. For including this information in the classification process, descriptive data must be represented using some scale. The data was represented as nominal data based on domain inputs. Some of the attributes processed are

- Additional Responsibilities
- Achievements
- Qualification

Qualification particularly may be significant in the evaluation process. It is generally mentioned as the degree obtained or pursuing. Based on the qualification details of the employee, qualification was also represented as category attribute. (Ranges are based on researcher's assumptions). For example PhD is treated as A, pursuing PhD as B and so on.

3.4 Supervised Learning – Classification

Classification is the supervised learning technique learns from the training objects. The main objective of the study is to classify the employees according to their performance category. Figure 2.2 describes the process of supervised leaning, which starts with data identification and concludes with evaluated classifier. This study also follows similar approach groups for building the classification model

The complexity of any DM task increases with the number of attributes being analyzed. There are several supervised as well as unsupervised dimensionality reduction techniques available. Attribute subset selection is very much crucial to a DM process and classification in particular. Choice of features must be made carefully because bad reduction may lead to a loss in the discrimination power and thereby a decrease in the accuracy of the resulting classifier.

The complexity of the data is very evident from the description of data itself (section 3.2) of the data itself. Since the attributes are distributed across hierarchies, and several attributes are repeated across hierarchies also, application of dimensionality reduction techniques to the entire dataset as a whole would have been inappropriate.

3.4.1 Elementary Prediction Model

After handling the missing data and required preprocessing of the data, first level classification considering all the parameters was performed. As mentioned in section 3.1, the approach for the performance evaluation is hierarchical, having different appraisal forms for different hierarchy. Basic classification was performed to understand dependencies among the parameters at different levels of hierarchy.

"Raw" classification without applying any dimensionality reduction technique was carried out on each of the above-mentioned datasets and also on the combination of the datasets. This elementary classification follows generic sequence of steps shown in the figure 3.4

Figure 3.4 Elementary Classification Model

Following different combinations of preprocessed datasets were used for this classification step-

- D1, D2, D3, D4 independently
- D1 &D2

- D1 &D3
- D1 &D 4
- D2 &D4
- D3 &D4
- D2 &D3
- D2 & D3 & D4

For these basic classifications, decision tree algorithm was the most suitable choice because of its interpretability. The tree structure generated as a result also provided important insights about significant attributes of the concerned dataset.

Basic Classification Results

Basic Classification 1

Data set – D4

Total No of attributes (Predictors) – 9

Data Type of attributes - Nominal

No of significant attributes (Attributes appearing in the decision tree) – 5

Classification Accuracy – 75.34%

Basic Classification 2

Data Set D1

Total No of attributes (Predictors) – 15

Data Type of attributes – Nominal and Numeric

No of significant attributes (Attributes appearing in the decision tree) – 6

Classification Accuracy – 49.31%

Basic Classification 3

Data Set D3

Total No of attributes (Predictors) – 50

Data Type of attributes - Nominal

No of significant attributes (Attributes appearing in the decision tree) – 10

Classification Accuracy – 52.03%

Basic Classification 4

Data Set D2

Total No of attributes (Predictors) – 59

Data Type of attributes - Nominal

No of significant attributes (Attributes appearing in the decision tree) – 13

Classification Accuracy – 45.02%

Basic Classification 5

Data Set D3 and D4

Total No of attributes (Predictors) – 59

No of significant attributes (Attributes appearing in the decision tree) – 7

Classification Accuracy – 80.82%

Basic Classification 6

Data Set D2 and D4

Total No of attributes (Predictors) – 68

No of significant attributes (Attributes appearing in the decision tree) – 12

Classification Accuracy – 71.23%

Basic Classification 7

Data Set D1 and D4

Total No of attributes (Predictors) – 24

Data Type of attributes – Nominal and Numeric

No of significant attributes (Attributes appearing in the decision tree) – 6

Classification Accuracy – 69.86%

Basic Classification 8

Data Set D2 and D3

Total No of attributes (Predictors) – 109

Data Type of attributes - Nominal

No of significant attributes (Attributes appearing in the decision tree) – 9

Classification Accuracy – 38.35%

Basic Classification 9

Data Set D1 and D2

Total No of attributes (Predictors) – 74

Data Type of attributes – Nominal and Numeric

No of significant attributes (Attributes appearing in the decision tree) – 12

Classification Accuracy – 47.97%

Basic Classification 10

Data Set D1 and D3

Total No of attributes (Predictors) – 65

Data Type of attributes – Nominal

No of significant attributes (Attributes appearing in the decision tree) – 15

Classification Accuracy – 36.98%

Basic Classification 11

Data Set D1,D2 and D3

Total No of attributes (Predictors) – 124

Data Type of attributes – Nominal and Numeric

No of significant attributes (Attributes appearing in the decision tree) – 14

Classification Accuracy – 43.83%

Basic Classification 12

Data Set D1, D2, D3 and D4

Total No of attributes (Predictors) – 133

Data Type of attributes – Nominal and Numeric

No of significant attributes (Attributes appearing in the decision tree) – 10

Classification Accuracy – 65.75%

Since the basic classification was performed for preliminary understanding of the data sets and their significance in the evaluation process, no attempt for improving the accuracy is made. Also the results do not highlight other performance parameters like ROC and other performance statistics.

The objective behind the classification using different datasets was to understand the significance of individual dataset to the performance category of an employee. It was found that the data set D4 gives the highest accuracy of classification. The reason for this very expected result is that the performance value is calculated directly from the attribute values of dataset D4. It was also observed that classification performed even with combined dataset without dataset 4 could not produce comparable results. This step thus highlights the dominance of this dataset in evaluating the performance. This led to the requirement of analyzing all parameters to find their relevance and contribution in the performance evaluation process without dataset D4. Thus the problem of classification was then redefined as finding mapping of attribute form remaining dataset onto the attributes of dataset D4. The insights obtained in this step are taken as input for subsequent phases of designing the final classifier.

3.4.2 Refined Classification Model

After the basic first level classification, it was revealed that the data set 4 is predominant and it controls the entire process of evaluation of a faculty member. The classifier thus would not consider attributes from this set as predictors, but they would be considered as class label attributes. The subset of significant attributes is determined using three different approaches integrated approach as follows -

3.4.2.1 Hierarchical Approach

As it is being highlighted in the previous sections, the nature of the appraisal process itself hierarchical. Thus, this approach exploits it to construct significant attribute set for the classifier. The steps of construction of the hierarchical classifier are as follows –

- Build classifier only for dataset 4
- Generate a set of class labels attributes - Each of the significant attribute obtained in step I is treated as a class label attribute.

- Perform classification for the each of the class label attributes considering all parameters in dataset 1, 2 & 3 as predictors.
- The union of significant attributes of the classifiers built in the previous step forms the set attributes to be used for final classification

Thus, the problem of single classification has been transformed to multiple classifications. Every classifier built would essentially follow following sequence of steps-

- Attribute subset selection
- Building the classifier
- Evaluate the classifier using the selected attributes
- Add the significant attributes to the final attribute subset

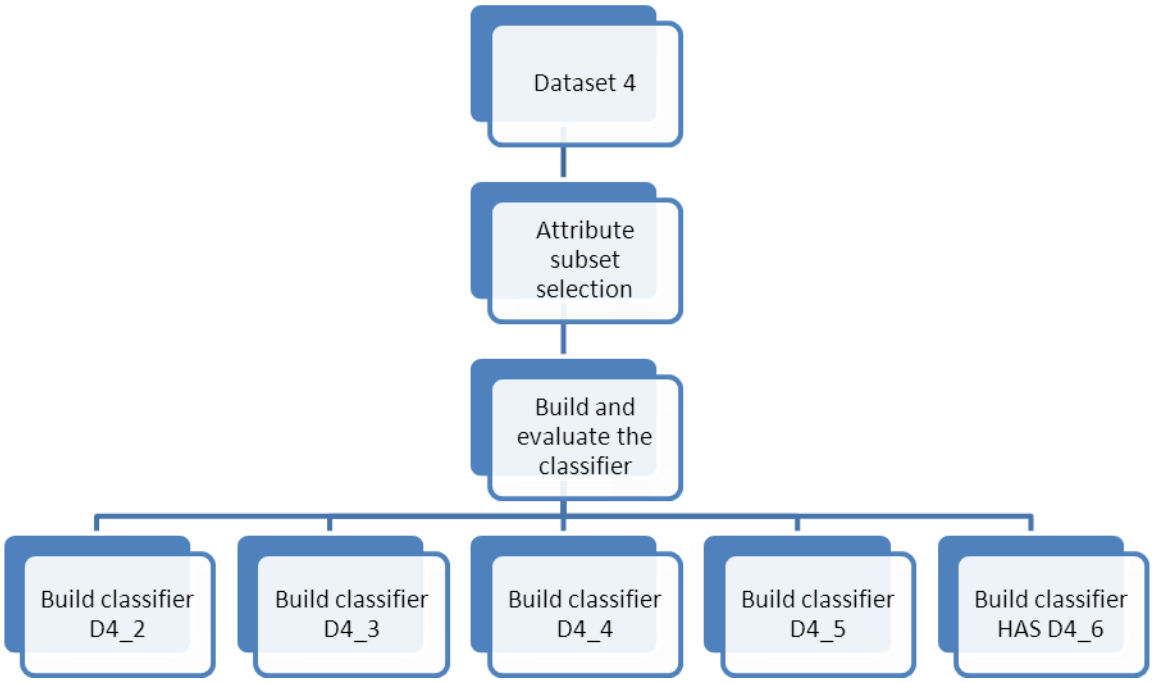


Figure 3.5 Hierarchical Classifier Approach

The process of hierarchical attribute construction thus starts with building the classifier for dataset 4 and continues with building a classifier for every significant attribute obtained in the classifier. Brief description of every classifier built in the process is given below-

Building classifier D4_M

No of significant attributes (Attributes appearing in the decision tree) – 5 represented as (D4_1, D4_2, D4_3, D4_4, D4_5, D4_6)

Out of the 9 attributes in the dataset only 6 are significant attributes. Attribute D4_1 is an independent variable (students rating). Thus for finding final attribute subset, each of these attributes are treated as class labels and a mapping of variables from form datasets D2, D2, & D3 to each of the above mentioned attributes needs to be determined. Thus every mapping results into a separate classification requirement and must follow the generic sequence of steps – attribute subset selection and classification.

Building classifier D4_2

Attribute D4_2 in the dataset 4 is the consolidated rating of an employee for his/her proficiency in executing the job as a teacher. Correspondingly the dataset 2 and 3 include sub parameters that are used to assess a teacher. Typically such attributes may include innovation in teaching, ability to answer students query, extra reading and information provided to students etc. These parameters are used as predictor variable to classify employees for the attribute D4_2 in the dataset 4. The dataset 3 and dataset 2 together contain 35 attributes which are requires to assess an employee for his role as teacher.

The attributes obtained in this step are added to a subset represented by name HAS_D4_2

Building Classifier D4_3

The attribute D4_3 is the assessment of the employee by first appraiser. This is the overall assessment of the faculty member for his role as a teacher, overall attitude, other potential that can be utilized by the organization there by resulting in the employee's growth etc. Thus this is the first appraiser's overall rating not directly stated but rather is based on the ratings given by him for all the other attributes in his appraisal form. Thus based on appraiser 1 responses, appraiser 2 gives the rating. The input thus is all the parameters of second appraiser evaluation form.

The attributes obtained in this step are added to a subset represented by name HAS_D4_3.

Building classifier D4_4

In consistent with other attributes of the dataset 4, attribute D4_4 also represents consolidated rating employees ability of working in the department. This attribute also defined by corresponding sub parameters in the dataset 2 and dataset 3. Typical of these sub parameters are attitude of the employee, sharing of the work etc. As in case of HAS_D4_2, this classification also considers task relevant attributes form the dataset 2 and 3.

As mentioned above the attributes from dataset 2 and dataset 3 are considered as input variables. Dataset D3 which basically is first appraiser's evaluation form contains his consolidated rating for D4_4. If sub parameters are replaced by consolidated rating, the accuracy of the classification improves.

The attributes obtained in this step are added to a subset represented by name HAS_D4_4

Building classifier D4_5

One section of the self-evaluation form contains general parameters about an employee's opinion about self-work asking for responses to attributes like factors influencing your performance, self-opinion of work etc. Attribute D45 is repeated in all three forms and is related to the employees work perspective in the context of organization as a whole and is reflected while executing the job. Thus relevant attributes from dataset 2 and 3 were used as input variables.

The attributes obtained in this step are added to a subset represented by name HAS_D4_5

Building classifier D4_6

While working in the capacity of a faculty member, it is expected to handle other allied responsibilities apart from classroom teaching. Attribute D46 in the dataset 4 is the consolidated rating of an employee for his/her proficiency in executing the other responsibilities along with the job of a teacher. Correspondingly the dataset 2 and 3 include sub parameters that are used to assess a teacher. Typically such attributes may include timely completion of the job assigned, follow up required etc. These parameters are used as predictor variable to classify employees for the attribute D4_2 in the dataset D4. The dataset D3 and dataset D2 together contain 20 attributes which are requires to assess an employee for his role as teacher. The summary of the classification process using the Bayesian classification algorithm is given below

The attributes obtained in this step are added to a subset represented by name HAS_D46

Final set of attributes HAS thus can be obtained as

HAS = D41 (The independent attribute) UNION {HAS_D42} UNION {HAS_D43} UNION {HAS_D44} UNION {HAS_D45} UNION {HAS_D46}

3.4.2.2 Direct Approach

This approach constructs significant attribute set with an objective of generating a mapping of attributes from datasets 1, 2 &3 directly to the performance category. The intermediate level of dependency of attributes with dataset 4 is not being considered. The sequence of steps followed to construct this subset is described below-

The set of attributes obtained after direct approach for attribute subset selection is represented as DAS. Initially, DAS ={}.

1. Integrate the datasets 1, 2 & 3
2. Apply information gain
3. If DAS ={} then add the attributes appearing in the decision tree to the attribute set DAS
4. If DAS <>{} then Remove the attributes not appearing in the decision tree from the DAS
5. Add the attributes with information gain value higher than attributes in DAS to the set DAS
6. Repeat step 2 to step 6 till no further improvement in the classification performance is possible.
7. Build the classifier using the subset obtained in the previous step.

The classifier approach is depicted in the figure 3.6

Figure 3.6 Direct Classifier Approach

3.4.2.3 Integrated approach

The final significant attribute subset is the union of subset obtained by hierarchical approach and sub set obtained by direct approach (Figure 3.7). The classification process then continues with building the classifier with the integrated sub set.

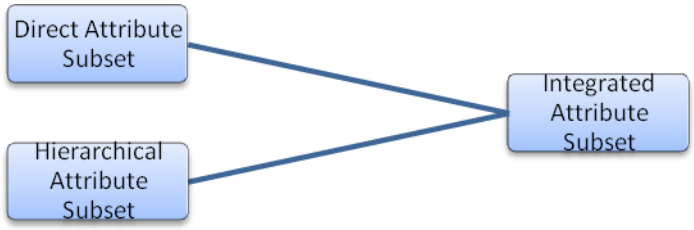


Figure 3.7 Integrated Approach for attribute Selection

The exploratory analysis need in the study has transformed a problem of single classification into multiple classifications. Decision tree and Bayesian classification algorithms are used for building the intermediate classifiers. Detailed discussion of the classifiers built in the study using above-mentioned algorithms and their performance is discussed in chapter 4.

3.5 Unsupervised Learning Approach

In the previous chapters, the supervised learning technique of classification was discussed in detail. Classification is an extremely important technique to for distinguishing group of objects but is based on previously known information in the form of training tuples. Clustering – An unsupervised learning technique is of great importance in a KDD process. It groups data into clusters based on attribute values describing the objects. It is very beneficial in understanding basic similarities among objects without requiring any extra input. As highlighted in the chapter 3, data was analyzed using unsupervised learning techniques to uncover similarities and dissimilarities present in among the tuples which are not captured in the classification process.

Even though there are a variety of clustering algorithms are available, the basic concept is the distance between two objects which decides how “similar” or “dissimilar” the two objects are. In

cluster analysis each tuple is considered as an object in n dimensional space where n is the number of parameters describing the tuple. Objects are grouped based on the distance between them which can calculated using different mathematical measures like Euclidian distance, Manhattan distance etc. Distance calculation depends on the type of the attribute (viz numeric, categorical, ordinal, and binary).

As mentioned in the literature review, several models have used clustering as feature selection measure. Since clustering is unsupervised learning technique, it is very efficient in uncovering basic interrelationships between objects. This information is then used to classify the objects using supervised learning methods. But the unlike the widely popular practice of clustering for feature selection for classification, clustering is used with a different perspective in this study. Employees are clustered with an objective of understanding the possibility of bias in the evaluation process across the organization.

The sequence of steps used to perform this unsupervised analysis is –

1. Normalize the data
2. Apply K-means clustering algorithm to the subgroups.
3. Find "cluster relevant" attributes

Supervised learning techniques used in the study are attributes oriented while the cluster analysis is tuples oriented. K-means clustering is a very well-known and commonly used clustering algorithm. It is a partitioning algorithm based on distance measures for similarity. Once the tuples are clustered, the clusters could be treated as classes and similarity criteria can be analyzed using supervised techniques.

K means is a popular partitioning based clustering algorithm. It is a distance-based technique that was originally designed for continuous data. Dissimilarity or distance calculation methods differ for different type of data like nominal, ordinal and binary. For using the in built K means MATLAB algorithm, the nominal data was converted to numeric category data. Thus, a simple mapping of (G S N) to (5 3 1) was applied to make the data suitable for application of the algorithm. Once data is represented using such scale and if the number of states is same for all variables, distance measures for interval scale variables can be applied while using K Means algorithm. Missing values are handled using knnimpute function, which replaces the missing

value with nearest neighbor. Since it is a distance-based technique, normalization of data is required. MATLAB code for normalization and K means clustering is given below-

Function Scale – Normalizes the data

```
function [ scaled ] = Scale(Data, Lower, Upper)
%UNTITLED2 Summary of this function goes here
%   Detailed explanation goes here

if (Lower > Upper)
   disp (['Wrong Lower or Upper values!']);
end
[MaxV, I]=max(Data);
[MinV, I]=min(Data);

[R,C]= size(Data);

scaled=(Data-ones(R,1)*MinV).*(ones(R,1)*((Upper-Lower)*ones(1,C)./(MaxV-
MinV)))+Lower;

end
```

Kmeans – Clusters the data based on the no of clusters given as input

```
    %clustering using k means algorithm
%Takes the name of the file as input and stores the result of clustering in
    %another file
    disp(['Enter the name of input file'])
    filename1=input('prompt','s');
    Data=xlsread(filename1);
    %A=ordinal(Data)
    result=knnimpute(Data);
    y=result;
    [r,c]=size(y);
    Lower=0;
    Upper=1;
    d=Scale(result,Lower,Upper);
    disp(['Enter the number of clusters (k}'])
    %k=input('prompt')
    size(d);
    [cidx] = kmeans(d,3);
    n=size(cidx);
```

```
figure(1)
for i=1:n
   index1=find(cidx==i);
   [um,un]=size(index1);
   if(um>0)
        disp(['Cluster-' int2str(i) ' = ' int2str(um)])
        y(index1,c+1)=i;
      plot(index1,i, 'r+')
   end
   hold on;
end
disp(['Enter the name of output file'])
filename2=input('promt','s');
xlswrite(filename2,y);
```

3.6 Classification Algorithms

Selection of appropriate algorithms for performing a DM task is very much crucial to the quality of the results obtained. An algorithm is selected based on the type and characteristics of the data. In chapter 2, an overview of different categories of classification algorithms is given. Since most of the attributes being considered in the study are categorical attributes, the most suitable classification algorithms are decision tree and Bayesian classification algorithm.

3.6.1 Decision Tree Classifiers

Decision tree is the most widely used classification algorithm. It remains one of the very popular classification algorithms even today because of its non-dependence on domain knowledge and comprehensibility.

A decision tree is a tree data structure consisting of decision nodes and leaves. A leaf specifies a class value. A decision node specifies a test over one of the predictive attributes, which is called the attribute selected at the node. For each possible outcome of the test, a child node is present. A test on a discrete attribute A has h possible outcomes A = d1, . . . , A = dh, where d1; : : : dh are the known values in domain(A). A test on a continuous attribute has 2 possible outcomes, A < t and A > t, where t is a threshold value determined at the node. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The algorithm, summarized as follows-

1. Create a node N;

2. If samples are all of the same class, C then
3. Return N as a leaf node labeled with the class C;
4. If attribute-list is empty then
5. Return N as a leaf node labeled with the most common class in samples;
6. Select test-attribute, the attribute among attribute-list with the highest information gain;
7. Label node N with test-attribute;
8. For each known value ai of test-attribute
9. Grow a branch from node N for the condition
test-attribute= ai;

10. Let si be the set of samples for which test-attribute= ai;
11. If si is empty then
12. Attach a leaf labeled with the most common class in samples;
13. Else attach the node returned by
Generate_decision_tree(si,attribute-list_test-attribute)

Thus, based on the information gain, a tree structure is constructed iteratively which best divides the training data. The procedure continues till there are tuples to divide and attributes to split.
J 48 – The C4.5 implementation of decision tree algorithm is used as a tool during the study.

3.6.2 Bayesian Classification

One of the simplest statistical classifier, Bayesian classifiers are used for categorical data. They predict the class membership probabilities. The naive Bayes classifier combines Bayesian probability model with a decision rule.

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector $\mathbf{X} = (x_1, x_2, \ldots, x_n)$
- Suppose there are *m* classes $C_1, C_2, \ldots, C_m$.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized

Where,

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

- If A_k is categorical, $P(x_k|C_i)$ is the # of tuples in C_i having value x_k for A_k divided by $|C_{i,D}|$ (# of tuples of C_i in D)

- If A_k is continous-valued, $P(x_k|C_i)$ is usually computed based on Gaussian distribution with a mean μ and standard deviation σ

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and $P(x_k|C_i)$ is

$$P(\mathbf{X}|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

A Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. This property of class conditional independence is a major limitation of Bayesian classifiers. In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers have worked quite well in many complex real-world situations.

3.6.3 Information Gain - Attribute Evaluation Measure

As mentioned in section 3.2, the data being analyzed is of categorical or nominal type. Also the numerical attributes in the original dataset were generalized to nominal attributes. Thus information gain is the natural choice for attribute evaluation criteria. Information gain is one of the most commonly used effective attribute information evaluation measure for categorical and nominal data. These measures typically determine the information gain from a feature. The information

gain from a feature X is defined as the difference between the prior uncertainty and expected posterior uncertainty using X. Feature X is preferred to feature Y if the information gain from feature X is greater than that from feature [22].

Information gain is an entropy based attribute evaluation measure which calculates the strength of an attribute in classifying a tuple based on following formula –

Info(D)= -∑pi * log2(pi)

Where pi is the probability that a tuple belongs to class Ci

Pi = |Ci|/|D|

For a discrete valued attribute A, based on the number of distinct values of the attribute information still required to classify could be obtained.($INFO_A(D)$. Gain(A) thus can be calculated as

Gain(A)= Info(D) - ($INFO_A(D)$

The attribute with maximum gain value is considered to be the best choice for splitting.

3.7 Tools Used

The study has used Weka 3.6.0 and MATLAB as analysis tools. Supervised learning approach was studied using Weka because of the rich classification algorithms. Since MATLAB allows more flexible treatment of individual tuples, unsupervised learning was studied in MATLAB.

Weka 3.6.0

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Advantages of Weka include:

- Freely availability under the GNU General Public License
- Portability: because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
- A comprehensive collection of data preprocessing and modeling techniques
- It provides many different algorithms for data mining and machine learning

- It is easily useable by people who are not data mining specialists
- It provides flexible facilities for scripting experiments
- it has kept up-to-date, with new algorithms being added as they appear in the research literature

MATLAB –

MATLAB ("MATrix LABoratory") is a tool for numerical computation and visualization. MATLAB® is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. Using the MATLAB product, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran. The basic data element is a matrix, so if you need a program that manipulates array-based data it is generally fast to write and run in MATLAB. MATLAB is a widely used tool in the scientific research.

It can be used for simple mathematical manipulations with matrices, for understanding and teaching basic mathematical and engineering concepts, and even for studying and simulating. It has an extended graphical function set which is very beneficial in curve plotting. The original concept of a small and handy tool has evolved to become scientific analysis workhorse. It is now accepted that MATLAB and its numerous Toolboxes can replace and/or enhance the usage of traditional simulation tools for advanced engineering applications. Its specialized toolboxes can be used for problem specific analysis tasks.
Some of the important capabilities of MATLAB include

- Developing Algorithms and Applications
- Analyzing and Accessing Data
- Visualizing Data
- Performing Numeric Computation
- Publishing Results and Deploying Applications

4. Results and Discussion

The study entitled “Employee performance prediction model” is aimed at application of DM techniques to develop a decision support tool for HRM to facilitate decision making. A multi process, multi-level and complex analysis is carried out to build a suitable classification model. To deal with of the complexity of the data, a thorough and effective KDD model is built. After a detailed discussion of the different techniques and the methods used in every phase of the research in preceding chapters, this chapter is a discussion of the results obtained in various sub processes and the evaluation of results.

Even though every KDD process essentially follows sequence of steps depicted in figure 1.3, the complexity of data analyzed during the study led to a KDD sub process in every step of the study. Thus the nature of the methodology could be described as “recursive”.

4.1 Classifier Evaluation

Classification the supervised learning process generates a classifier model after applying data preprocessing, attribute subset selection and classification algorithm. The classifier is essentially a set of rules to be used to classify a new instance. Before using the classifier for unseen data it must be evaluated. There are two widely used classifier evaluation techniques

- **Cross Fold Validation –**

In K cross fold validation, the initial data are randomly partitioned into k mutually exclusive subsets or folds, D1,D2,…….,Dk, approximately of same size. Training and testing is performed iteratively K times. In iteration I, Di is reserved as test set and remaining subsets are considered as training sets. The process continues in a similar way for other datasets also.

Leave one out and stratified cross validation are special cases of cross validation.

10 fold cross validation is recommended for evaluation of a classifier because of its low variance and bias.

Other important parameters that should evaluated are

- **Confusion matrix** is a m by m matrix where m is the number of classes is a very useful tool for analyzing the prediction accuracy of the classifier. For a good classifier, most of the tuples would represent across the diagonal.

- **ROC** – Receiver operating characteristics curves are visual representation of classifier accuracy and generally used to compare two classifiers. It essentially a plot of true positive rate and false positive rate of prediction by the classifier. ROC curve for a less accurate model will be closer to the diagonal line indicating more false positives. In a good model, true positive rate will be more and the curve moves steeply up from zero.

The following sections present the performance details for every classifier built in the course of the study.

4.1.1 Elementary Classifier

Elementary classifiers without any dimensionality reduction were constructed for various datasets. The performance details of two classifiers are presented again in this section for comparison.

Classifier 1 – For dataset D1, D2 and D3

Data Set D1,D2 and D3

Total No of attributes (Predictors)	:	124
Data Type of attributes	:	Nominal
No of significant attributes (Attributes appearing in the decision tree)	:	14
Classification Algorithm Used	:	Tree – J48
Classification Accuracy	:	43.83%

Classifier 2 – Only dataset D4

The dataset considered for this evaluation is dataset 4

Number of Parameters	:	9

Class Label Attribute	:	CLASS (Values A, B, C)
Type of Parameters	:	Nominal
Attribute Selection Algorithm Used	:	None
Classification Algorithm Used	:	Decision Tree - J48
Accuracy	:	75.34%

The classification was performed for 9 nominal parameters of dataset4. Since this is basic classification, no attribute subset selection measure has been used. The accuracy of classification is significant 75.35% .The tree structure constructed shows only 6 parameters. As it has been emphasized in the chapter3 the accuracy of the classifier for dataset D4 is not comparable with classifiers with other dataset or combinations of other datasets. Since these classifiers are very basic in nature other performance parameters like ROC, confusion matrix are not discussed.

4.1.2 Intermediate Classifiers

Six intermediate classifiers are constructed in the process of building the hierarchical classifier. In this section a detailed performance statistics for each of the intermediate classifier is presented.

Classifier D4_ M

The dataset considered for this evaluation is dataset 4

Number of Parameters	:	9
Class Label Attribute	:	CLASS (A, B ,C)
Type of Parameters	:	Nominal
No of Significant attributes	:	6
Attribute Subset Selection Method Used	:	Decision Tree, Chi squared evaluation
Classification Algorithm Used	:	Naïve Bayes

Classifier Representation

The Naïve Bayes classifier maximizes

P(X|Ci) P(Ci)

Where,

X, the tuple to be classified is described as

X= (D4_1, D4_2, D4_3, D4_4, D4_5, D4_6)

The classifier calculates P(X|Ci) based training set

P(Ci), Prior probabilities of each class

Class B: P(C) = 0.44736842

Class A: P(C) = 0.26315789

Class C: P(C) = 0.28947368

Accuracy 84.93%

Table 4.1 Classifier Performance Parameters

Accuracy 84.93%

Class	TP	FP	Precision	Recall	ROC
Class A	0.737	0.019	0.933	0.879	0.982
Class B	0.030	0.225	0.775	0.824	0.948
Class C	0.81	0.019	0.944	0.872	0.991

Confusion Matrix

	A	**B**	**C**
A	**48**	**3**	**0**
B	**9**	**4**	**0**
C	**7**	**1**	**1**

ROC Curves

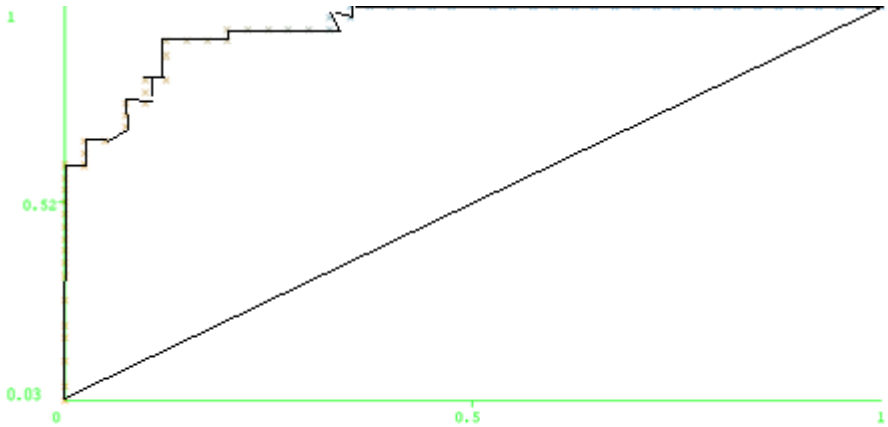


Figure 4.1 Class B Classifier D4_M

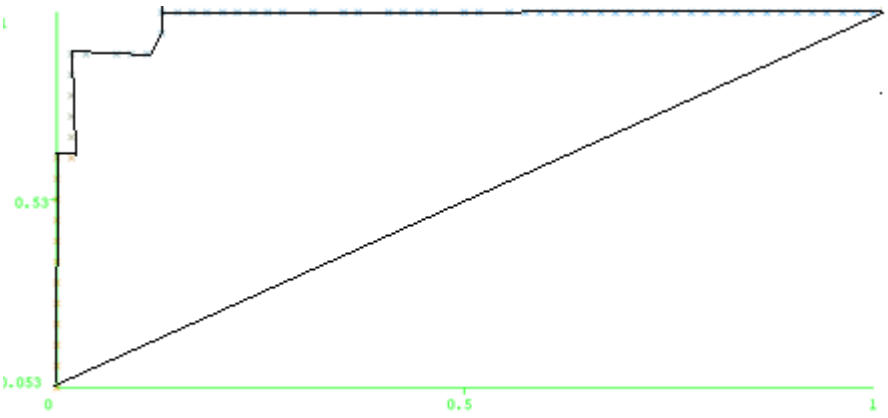


Figure 4.2 Class A Classifier D4_M

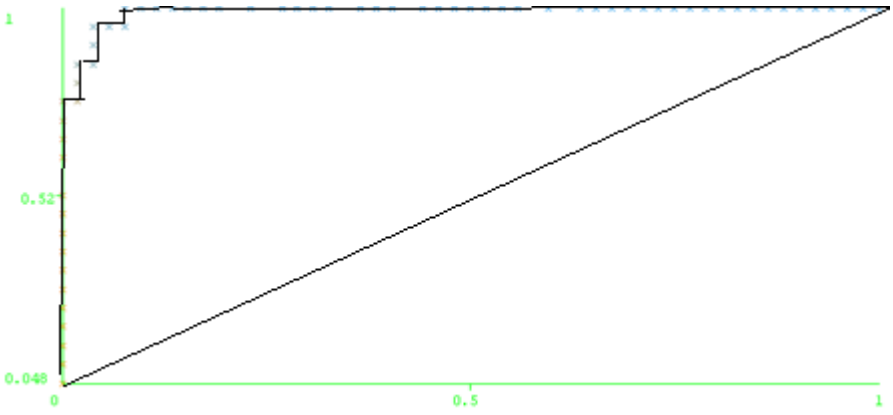


Figure 4.3 Class C Classifier D4_M

The classification model for dataset 4 is refi**nement** of the raw classification performed in the previous basic classification. The attributes appearing in the decision tree of basic classification were considered as the significant parameters. The classification was performed using Naïve Bayes algorithm. The accuracy of the classification performed is 84.9% which is very good. Since the dataset 4 represents consolidated ratings, the attributes are inherently independent of each other. The assumption of conditional independence of Naïve Bayes holds good to get significant performance of classification. Also, the classifier performs extremely well for classifying tuples with class B as compare to tuples with class A and C.

Classifier D4_2

Input dataset	:	Related parameters form D2 and D3
Total No of Attributes	:	35
Type of attributes	:	Nominal
Attribute selection Measure	:	Decision Tree, Information Gain
No of Significant Attributes	:	11
Classification Algorithm Used	:	Decision Tree J 48

Table 4.2 Classifier Performance Parameters

Accuracy 71.23%

Class	TP	FP	Precision	Recall	ROC
Class G	0.69	0.159	0.741	0.714	0.797
Class S	0.813	0.244	0.722	0.765	0.75
Class N	0.5	0.066	0.6	0.861	0.681

Confusion Matrix

	S	**G**	**N**
S	**20**	**7**	**2**
G	**4**	**26**	**2**
N	**3**	**3**	**6**

The training dataset for this classifier is task relevant attributes form dataset 2 and dataset 3. Initial 35 parameters were reduced to a set of 11 after application of iterative information gain and decision tree. Since the task of classification involves sub parameters, the possibility of interdependence between the attributes is high. The decision tree algorithm used for classification gives reasonable accuracy of 71.23%. Class II prediction accuracy is higher as compared to other classes. Also for class III, though true positive rate is significantly low, since false positive rate is also very low, the ROC statistics is better than other two classes.

Classifier D4_3

Input dataset : D3

Total No of Attributes : 50

Type of attributes : Nominal

Attribute selection Measure : Information Gain

No of Significant Attributes : 4

Classification Algorithm Used : Naive Bayes

Classifier Representation

The Naïve Bayes classifier maximizes

P(X|Ci) P(Ci)

Where,

X, the tuple to be classified is described as

X= (D3_1, D3_2, D4_46, D4_50)

The classifier calculates P(X|Ci) based training set

P(Ci), Prior probabilities of each class

Class G: P(C) = 0.37

Class N: P(C) = 0.08

Class S: P(C) = 0.55

Table 4.3 Classifier Performance Parameters

Accuracy 69.8%

Class	TP	FP	Precision	Recall	ROC
Class G	0.66	0.174	0.68	0.63	0.797
Class N	0.8	0.044	0.571	0.8	0.75
Class S	0.732	0.344	0.732	0.732	0.681

Confusion Matrix

	G	**N**	**S**
G	**17**	**0**	**10**
N	**0**	**4**	**1**
S	**8**	**3**	**30**

The classifier model takes all the parameters from dataset 3 to find appropriate mapping rule. After the feature extraction, the parameters obtained represent group level attributes and not the sub parameters. These attributes thus are independent of each other and the Bayesian classifier constructed gives the accuracy of 69.8%. The confusion matrix also highlights the significant miss classification between the classes G and S. The tuples falsely classified as S when they

belong to G and vice versa is very high. For the class N the prediction statistics is adequately good.

Classifier D4_4

Input dataset	:	Related Parameters from D2 and D3
Total No of Attributes	:	20
Type of attributes	:	Nominal
Attribute selection Measure	:	Decision Tree
No of Significant Attributes	:	3
Classification Algorithm Used	:	Naive Bayes

Classifier Representation

P(X|Ci) P(Ci)

Where,

X, the tuple to be classified is described as

X= (D3_3, D2_34, D2_38)

The classifier calculates P(X|Ci) based training set

P(Ci), Prior probabilities of each class

Class G: P(C) = 0.25

Class N: P(C) = 0.08

Class S: P(C) = 0.67

Table 4.4 Classifier Performance Parameters

Accuracy 72.6%

Class	TP	FP	Precision	Recall	ROC
Class S	0.96	0.783	0.727	0.96	0.677
Class G	0.278	0.018	0.833	0.417	0.743
Class N	0	0.015	0.0	0	0.509

Confusion Matrix

	S	**G**	**N**
S	**48**	**1**	**1**
G	**13**	**5**	**0**
N	**5**	**0**	**0**

The Bayesian classifier constructed gives the accuracy of 72.6% which is reasonable. But the classifier false positive rate is significantly high which hampers the overall performance in correct prediction of tuple classes. From the confusion matrix, it can be observed that the prediction accuracy for class G is very good but class G has significantly low true positive rate. Also class N was never predicted correctly.

Classifier D4_5

Input dataset : Related Parameters of D3 and D3

Total No of Attributes : 9

Type of attributes : Nominal

Attribute selection Measure : Decision Tree

No of Significant Attributes : 3

Classification Algorithm Used : Naive Bayes

Classifier Representation

P(X|Ci) P(Ci)

Where,

X, the tuple to be classified is described as

X= (D2_55, D3_26, D3_50)

The classifier calculates P(X|Ci) based training set

P(Ci), Prior probabilities of each class

Class G: P(C) = 0.18

Class N: P(C) = 0.13

Class S: P(C) = 0.68

Table 4.5 Classifier Performance Parameters

Accuracy 72.6%

Class	TP	FP	Precision	Recall	ROC
Class G	0.941	0.727	0.75	0.835	0.712
Class G	0.308	0.067	0.5	0.308	0.764
Class N	0.111	0	0.111	0.2	0.575

Confusion Matrix

	S	**G**	**N**
S	**48**	**3**	**0**
G	**9**	**4**	**0**
N	**7**	**1**	**1**

The classifier constructed using Naïve Bayes thus gives the prediction accuracy of 72.06%. But as with the previous classifiers, the prediction accuracy for class S is very good as compared to classes G and N.

Classifier 6

Input Data : Related data from dataset 2 & 3

Total No of attributes (Predictors) : 10

Data Type of attributes : Nominal

Class Label Attribute : D4_5

Attribute subset selection measure used : Information Gain

No of significant attributes : 3

Classification Algorithm Used : Naïve Bayes

Table 4.6 Classifier Performance Parameters

Accuracy 67.12%

Class	TP	FP	Precision	Recall	ROC
Class S	1	0.889	1	0.793	0.608
Class G	0.063	0	1	0.118	0.622
Class N	0.182	0	0.182	0.308	0.516

Confusion Matrix

	S	**G**	**N**
S	**46**	**0**	**0**
G	**15**	**1**	**0**
N	**9**	**0**	**2**

As it can be observed from the confusion matrix, the statistics for class S is better than other two classes. Even though the classifier gives overall accuracy of over 72%, the false positive rate for classes N and G are very large.

4.1.3 Hierarchical Classifier

Each of the intermediate classifiers built in the process also contribute to the feature extraction process. The union of significant attributes of all the intermediate classifiers is considered as initial attribute set for the hierarchical classifier, which is further, refined in the process.

Input Data : As mentioned above

Total No of attributes (Predictors) : 24 + 1 (D4_1)

Data Type of attributes : Nominal

Class Label Attribute	:	CLASS
Attribute subset selection measure used	:	Information Gain
No of significant attributes	:	7
Classification Algorithm Used	:	Decision Tree J48

Table 4.7 Classifier Performance Parameters

Accuracy 79.45%

Class	TP	FP	Precision	Recall	ROC
Class B	0.879	0.275	0.725	0.879	0.796
Class A	0.632	0.037	0.857	0.727	0.793
Class C	0.81	0.038	0.895	085	0.875

Confusion Matrix

	B	**A**	**C**
B	**29**	**2**	**2**
A	**7**	**12**	**0**
C	**4**	**0**	**17**

ROC Curves

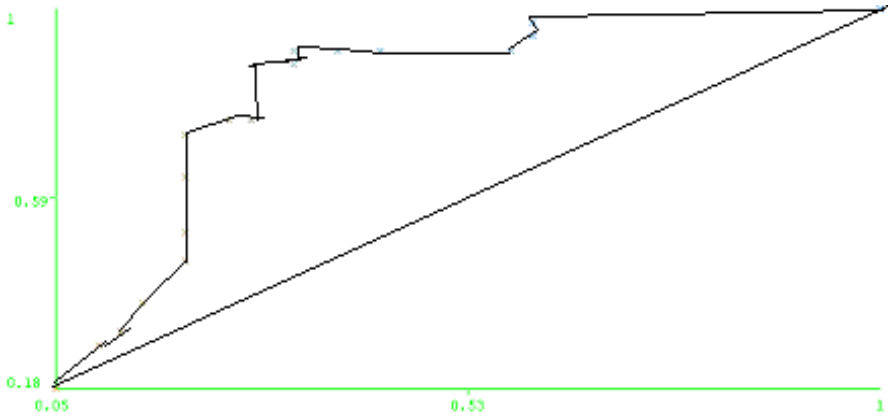


Figure 4.4 Class B Classifier Hierarchical

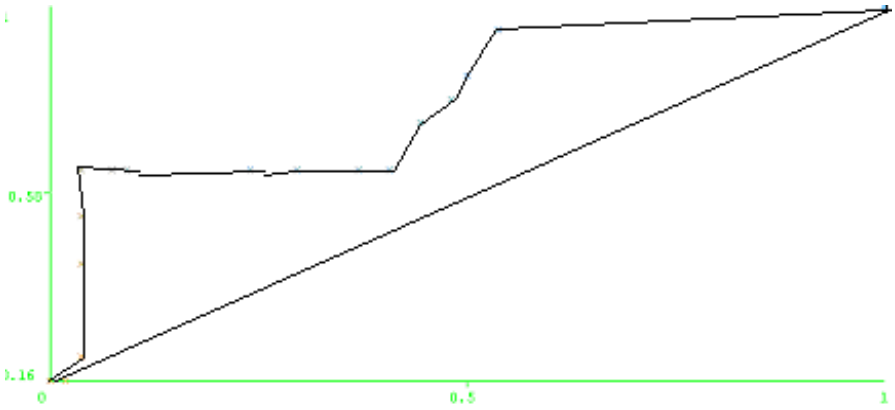


Figure 4.5 Class A Classifier Hierarchical

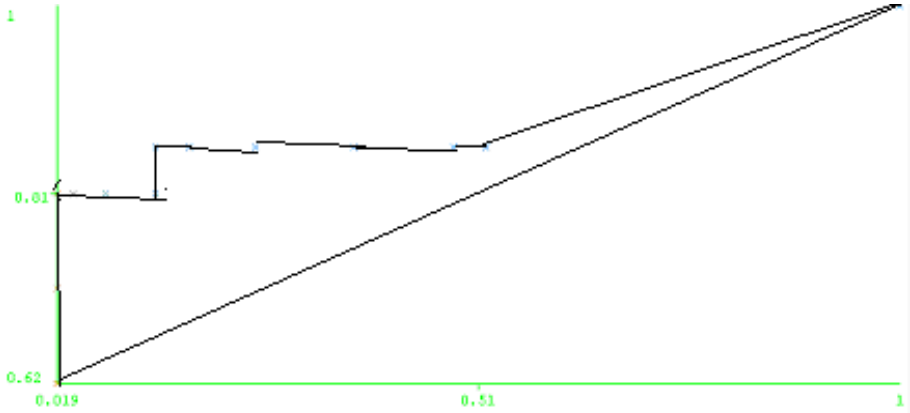


Figure 4.6 Class C Classifier Hierarchical

The tree classifier constructed using tree J48 has seven significant parameters. The overall accuracy of prediction is 79.45, which is very good. Also the confusion matrix shows high rate of true positives. But the mis classification rate for classes A and C being classified as B is high. ROC curve for class B is significantly better than class A and Class C.

4.1.4 Direct Classifier

Through Direct classifier an attempt of mapping of parameters directly to performance category has been made.

Input Data	:	As mentioned above
Total No of attributes (Predictors)	:	124
Data Type of attributes	:	Nominal
Class Label Attribute	:	CLASS
Attribute subset selection measure used	:	Information Gain and tree
No of significant attributes	:	7
Classification Algorithm Used	:	Decision Tree J48

Table 4.8 Classifier Performance Parameters

Accuracy 76.71%

Class	TP	FP	Precision	Recall	ROC
Class B	0.909	0.25	0.75	0.909	0.858
Class A	0.684	0.111	0.684	0.684	0.844
Class C	0.619	0.019	0.929	0.619	0.758

Confusion Matrix

	B	**A**	**C**
B	**30**	**2**	**1**
A	**6**	**13**	**0**
C	**4**	**4**	**13**

ROC

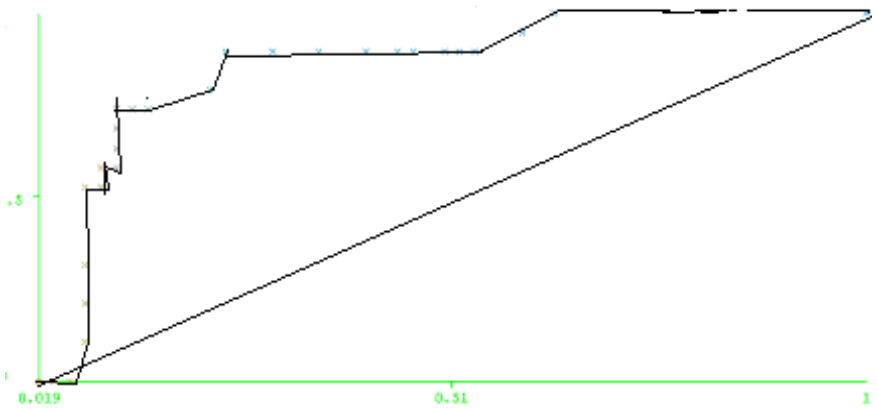


Figure 4.7 Class A Classifier Direct

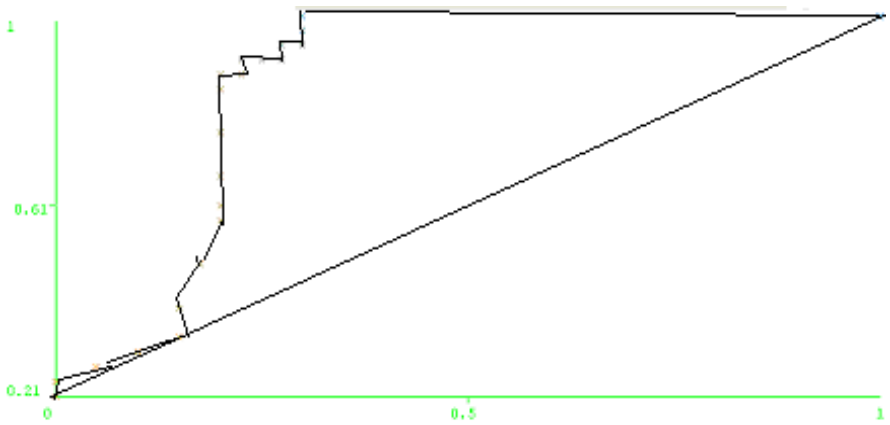


Figure 4.8 Class B Classifier Direct

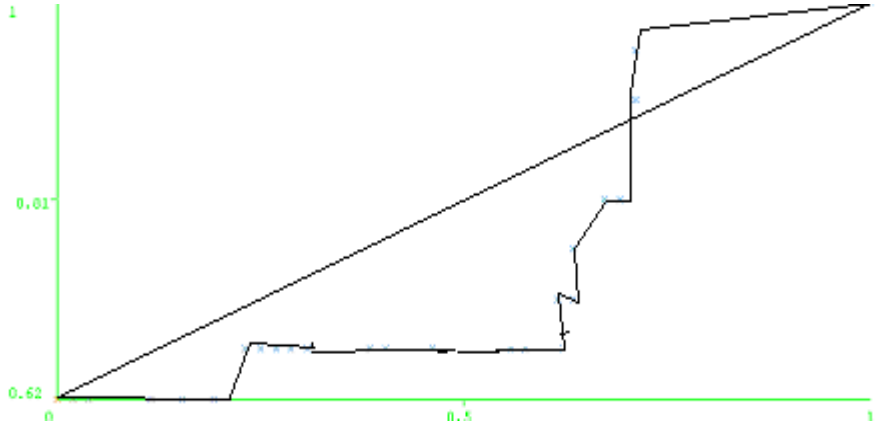


Figure 4.9 Class C Classifier Direct

The direct classifier starts with all 124 parameters and with iterative information gain and decision tree approach, selects 7 attributes as significant. With these 7 attributes, the classfier gives the accuracy of 76.71% which very good. The basic classification accuracy for the same dataset is below 50%. Performance statistics for class A and class B is reasonable. But the class C statistics is not appreciable because of high mis classification rates.

4.1.5 Integrated Classifier

Integrated classifier integrates the significant attributes obtained from direct as well as hierarchical approaches.

Input Data	:	As mentioned above
Total No of attributes (Predictors)	:	20
Data Type of attributes	:	Nominal
Class Label Attribute	:	CLASS
Attribute subset selection measure used	:	Information Gain and tree
No of significant attributes	:	7
Classification Algorithm Used	:	Decision Tree J48

Table 4.9 Classifier Performance Parameters

Accuracy 75.34%

Class	TP	FP	Precision	Recall	ROC
Class B	0.909	0.275	0.732	0.909	0.851
Class A	0.632	0.13	0.632	0.632	0.769
Class C	0.619	0.0	0.619	0.765	0.822

Confusion Matrix

	B	A	C
B	**30**	**3**	**0**
A	**7**	**12**	**0**
C	**4**	**4**	**13**

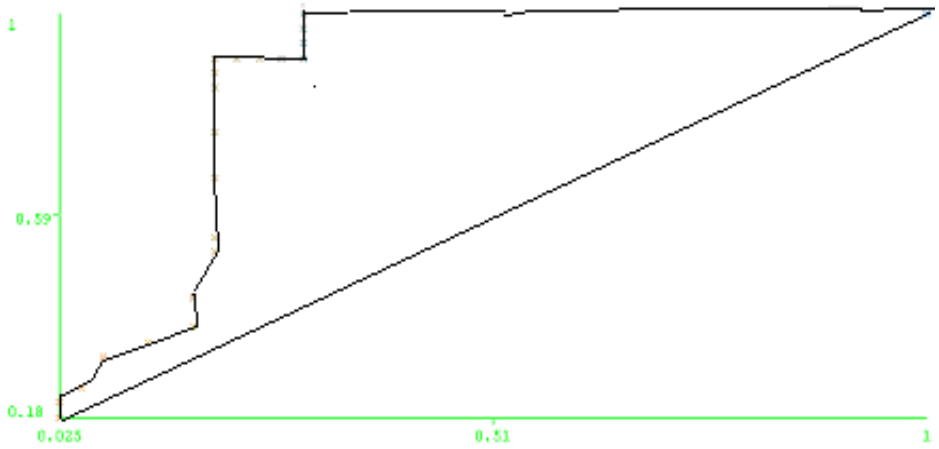


Figure 4.10 Class B Classifier Integrated

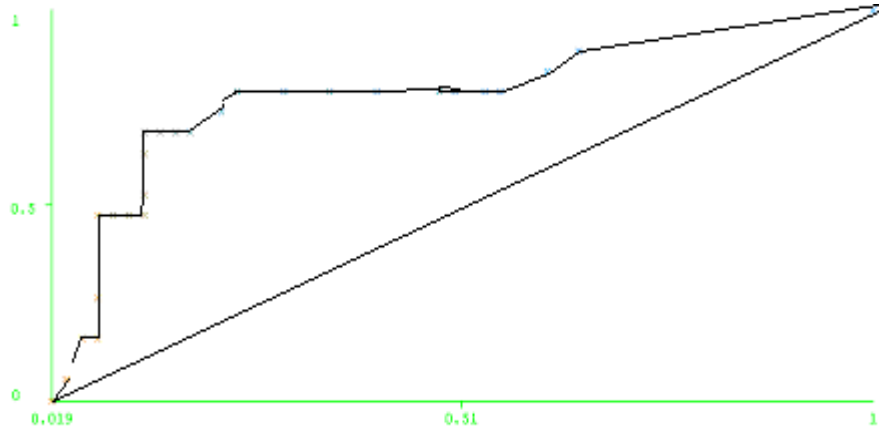


Figure 4.11 Class A Classifier Integrated

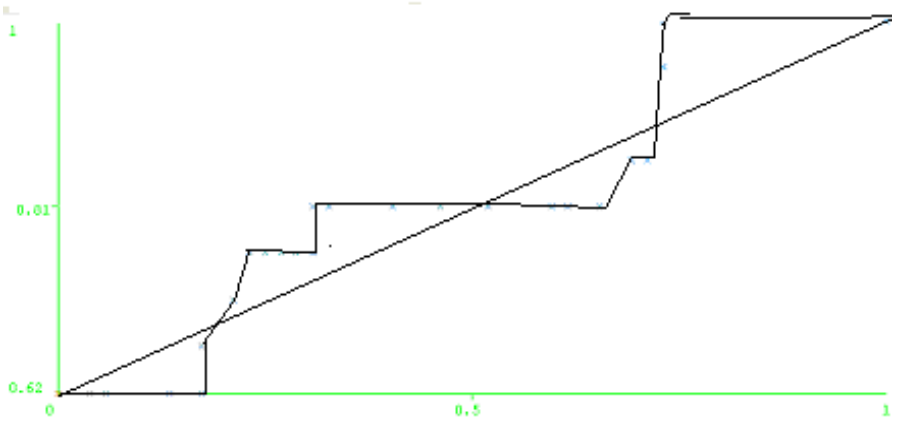


Figure 4.12 Class C Classifier Integrated

The integrated classifier is constructed for the integrated subset. (Combination of hierarchical and direct). The accuracy of the classifier is above 75%. This classifier also has low performance statistics for class C as compared to classes A and B.

4.2 Cluster Analysis

Cluster analysis is performed to analyze the natural similarities among the objects. Following figures present assignment of elements to individual clusters and silhouette representation of clusters to study inter relationships between the clusters.

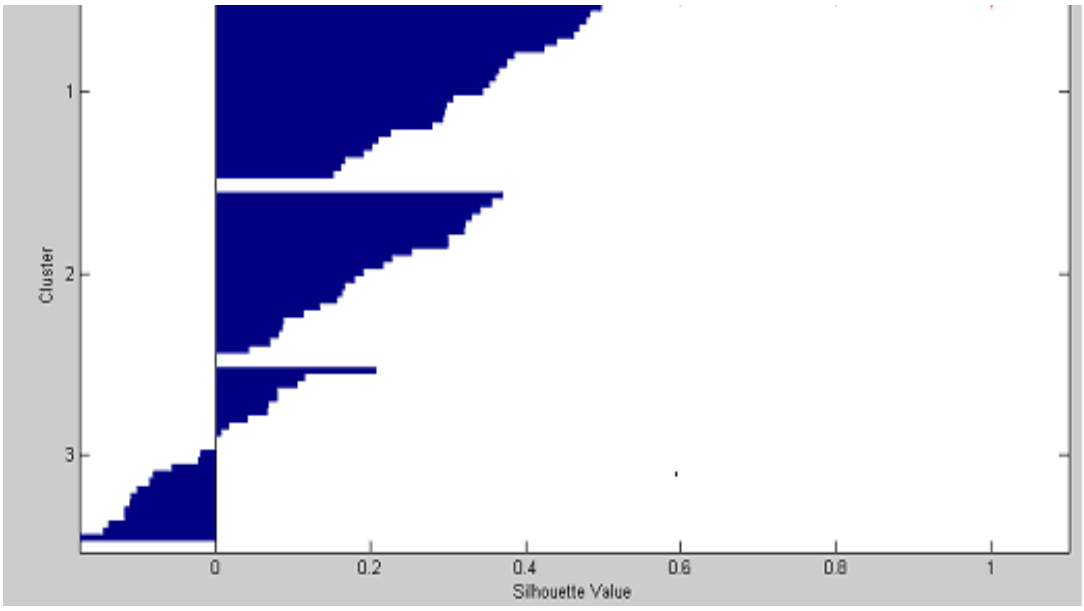


Figure 4.13 Cluster Silhoutte Plot

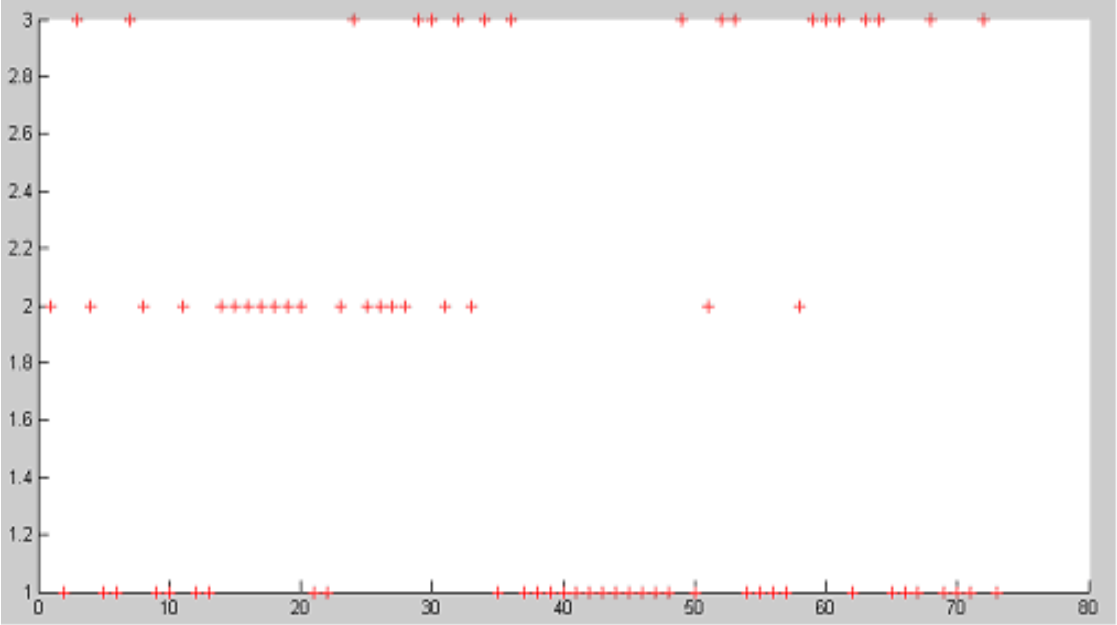


Figure 4.14 Assignment of Objects to Clusters

4.3 Discussion

Important findings of the study and interpretations of the results obtained are summarized in this section.

- Total 9 (6 intermediate and 3 final classifiers) classifiers were constructed during the course of the study. Additionally elementary classification was performed on 12 different pre processed datasets.

- The intermediate classifiers constructed in the process generate the significant attribute set for the hierarchical classifier. The first in the category of intermediate classifier, D4_M, is the classifier for dataset D4 and gives a very good accuracy of 84.9%. The dominance of this dataset has already been discussed. The remaining classifiers of this category give the prediction accuracy in the range of 68% to 73%. This range even though indicates a reasonable accuracy but also emphasizes on requirement for improvement. It also highlights element of subjectivity in the evaluation process. All these classifiers give a good performance for class S (S signifies the average rating), but the classifier does not

give satisfactory results for G (above average) and N (Below average) categories. This highlights the need for more clarity in the evaluation of extreme cases. N (below average) class in particular has a very poor performance statistics.

- The hierarchical classifier is constructed after all intermediate classifiers and gives a considerably good accuracy rate of 79.4%. The ROC curve for class B is (figure 4.4) indicates a balanced performance of the classifier for the class B. Class A performance parameters are also comparable with a reasonable ROC area. The performance of the classifier in predicting Class C employees is not adequate. As mentioned, the factors influencing the classifier performance could be the variation in the evaluation criterion from one appraiser to the other (including self evaluation). Also the prediction error gets propagated from the intermediate classifiers to the final hierarchical classifier. But it must also be noticed that even with the accuracy rage of 68% - 73% for intermediate classifier, the final classifier could achieve the accuracy of 79.4% emphasizing on the effectiveness of the attribute selection process.

- The direct classifier does not consider the intermediate level of dependency and directly finds a mapping for the final performance category. Even though the prediction accuracy is comparable with the hierarchical classifier, the prediction capability for the class C is very low with a poor ROC (figure 4.9) curve. The number of significant attributes is reduced from 124 to 7 with an iterative application of information gain and decision tree. But the attribute set appears to be biased more towards the attributes reflecting classroom teaching proficiency of the employee. The possibility of the classifier being sensitive to the training data set cannot be ruled out.

- The integrated classifier gives the prediction accuracy of over 75%, which is comparable with direct and hierarchical classifier. The performance statistics is very much similar to the hierarchical classifier but the limitation with class C performance is prorogated from direct to integrated classifier. This indicates the effectiveness of hierarchical attributes over the direct subset in the process of employee performance classification.

Cluster Analysis Results

Cluster analysis, the unsupervised learning approach was applied to understand the natural similarities among the employees. The result of cluster analysis indicates that

- The assignment of individual objects to clusters is different from class assignment initiating a requirement for further investigation.
- The silhouette plot of cluster in Figure 4.13 indicates weak inter cluster boundaries.
- It could be inferred from observations that the natural grouping of the elements is different from the "supervised" groupings.
- Unsupervised learning results could be analyzed to understand inherent similarities among employee performances and to understand the effect of bias in the evaluation process.

5. Summary and Conclusion

Data mining, with its effective knowledge acquisition techniques, is at the heart of any intelligent analysis. The study was performed with an objective of knowledge discovery in the domain of HRM by applying DM techniques. Data from academic industry was considered for the analysis. This chapter summarizes the work done during the course of the study along with the limitations and challenges and concludes with future scope of work.

The employee performance prediction model constructed after an exhaustive analysis of the data has been successful in its objectives. The study has used supervised and unsupervised learning approach to achieve the objectives. Though supervised learning approach was emphasized more, the findings of the unsupervised learning lead to a new analysis requirement and a subsequent thorough investigation of data with a new perspective. The evaluations of the classifiers built in the process justify the effectiveness of DM techniques in the knowledge discovery in the domain of HRM.

5.1 Summary

The study was carried out in five steps – data collection, data preprocessing, dimensionality reduction, classifier construction and classifier evaluation. Highlights of the study can be mentioned as

1. The hierarchical characteristics of the data worked as the driving factor for many sub processes employed by the study.
2. Correlation analysis though did not reveal redundancies but could provide insights into basic interrelationships between various parameters at different levels. Some of the significant correlations can be mentioned as-
    - Appraiser 1 and appraiser 2 ratings are more strongly correlated than employee self-evaluation ratings.
    - Students' evaluation of the teacher is highly correlated with the performance of the teacher.

- Writing of the faculty work diary is the only attribute which is negatively correlated with the performance.

3. Dimensionality reduction was the most complicated task because of the huge number of attributes. Thus, three different approaches – direct, hierarchical and integrated were employed in the study to capture the complexities of the data. The number of attributes was reduced to 7 from initial 124 in the process.

4. Three classifiers were constructed – hierarchical, direct and integrated. The accuracy of the three classifiers is comparable. But the hierarchical classifier gives better prediction performance as compared to the other two classifiers.

5. Prediction statistics suggest high accuracy rates for class B as compared to class A and class C.

6. Cluster analysis indicates different grouping of the employees that needs to be analyzed.

Factors influencing the prediction accuracy could be mentioned as

1. Subjectivity in the evaluation

2. Limitation of the three point nominal scale used for ratings. Particularly, such inadequate rating scale has negative effect and may result in wrong prediction along at the class boundaries (A classified as B or B classified as A).

3. The mapping of three point nominal scale onto a five point numeric scale by the top level hierarchy evaluation also affects the accuracy of prediction.

4. 45 parameters are repeated - they appear in the self as well as appraiser 1 evaluation. (together 90). This brings in lot of redundancy and complexity in the prediction process and affects the results of prediction performance.

5.2 Limitations and Challenges

In the previous section a brief overview of the findings of the study was presented. This section discusses limitations of the study and challenges faced during the course of the study.

- In any organization, employee performance data is strictly kept confidential. Since this data is about the performance parameters, revealing it may have ethical and behavioral issues. Thus, though utmost transparency is expected in the evaluation process; the secrecy of evaluation details is strictly maintained. Getting sufficient data, thus, is the biggest challenge for any study of this type. This study also has to put up with this limitation.

- Minimum of two individuals are involved in the process of appraisal - the appraiser and the appraisee. Depending on the organizational structure, the no of appraisers varies. Additionally appraisal parameters tend to be qualitative. Thus capturing the qualitative rating on a nominal or numeric scale poses difficulty in correct rating. Also, the element of subjectivity in an appraiser rating cannot be completely eliminated. These issues can be partially resolved using weightage factors for parameters and appropriate rating scales but cannot be eliminated entirely. Generally the performances are "normalized" by the topmost hierarchy to minimize the effect of bias. This bias if analyzed may provide important insights in the process of performance evaluation.

5.3 Future Work

The employee performance prediction model constructed in this study is can further be extended and several different techniques can be applied to the appraisal data. This section presents a brief discussion on possible future enhancements.

- **Predicting Performance as a continuous value –** The study can be extended to predict the performance of employee as a continuous value instead of predicting performance category of the employee. A comparative analysis of the category prediction (Classification) model and a value prediction would help to choose a more robust model.

- **Application of other Classification Algorithms –** Classification and prediction is the most widely studied technique of DM. Because of the categorical nature of the parameters, the choice of algorithms is restricted to Decision tree and Bayesian classification. With appropriate transformation of data on a numeric scale, data can be analyzed using other algorithms like neural networks.

- **GroupWise Analysis to Understand the Variation in the Evaluation –** The data could be split into different logical groups like departments or deaneries. All the processes followed in the study could be applied to the groups separately. This would help in understanding the subjectivity in the evaluation process. (This analysis is subject to availability of required data).

- **Integration of supervised and unsupervised learning** - As it is highlighted in chapter 4, the unsupervised grouping of employees is different from the supervised grouping employees. The study thus could be continued to understand the natural similarities among the objects. Such analysis would essentially help in uncovering other interesting patterns. For example cluster 1 contains some elements of class A, some elements of class B and some elements of class C. Understanding these factors other than discovered in the classification process provide additional insights might be beneficial for certain HRM decisions like allocation of right people to right job. If the relationships discovered are strong enough, unsupervised results could further be integrated with the supervised results to enhance the original performance classification process. This integration may follow two different approaches. In the first approach, along with the performance class of the employee, cluster analysis also generates a set of class labels. Thus an employee belongs to two classes - the original class and the new classes obtained from cluster analysis. In the second approach, the original class label could be transformed to a new set of class labels to incorporate cluster results. Such integration of supervised and unsupervised learning would enhance the performance evaluation process.

- **Implementation** – The analysis could be implemented to develop a tool. This tool may act as a decision support tool for HR personnel.

Bibliography

1. Zhao Xin, "**A Study of Performance Evaluation of HRM: Based on Data Mining**", 978-0-7695-3480-0/08 $25.00 © 2008 IEEE ,DOI 10.1109/FITME.2008.133

2. Jayanthi Ranjan, D P Goyal, S I Ahson "**Data mining techniques for better decisions in human resource management systems**", **International Journal of Business Information Systems 2008 - Vol. 3, No.5 pp. 464 - 481**

3. Hamidah Jantan, Abdul Razak Hamdan, and Zulaiha Ali Othman, "**Knowledge Discovery Techniques for Talent Forecasting in Human Resource Applications**", World Academy of Science, Engineering and Technology 50 2009

4. Chen-Fu Chien, Li-Fei Chen, "**Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry**", Expert Systems with Applications Volume 34, Issue 1, January 2008, Pages 280-290

5. Han Jing ," **Application of Fuzzy Data Mining Algorithm in Performance Evaluation of Human Resource**", 2009 International Forum on Computer Science-Technology and Applications, 978-0-7695-3930-0/09 $26.00 © 2009 IEEE,

6. *Jing Han*1, *Yanzhi Dong*2, "**Application of Association Rules Based on Rough Set in Human Resource Management**" , The Sixth Wuhan International Conference on E-Business - Innovation Management Track……2407

7. Adem Karahoca, Dilek Karahoca, Osam Kaya, "**Data Mining To Cluster Human Performance bu using Online Self Clustering Method**", 1st WSEAS International

Conference on Multivariate Analysis and its Application in Science and Engineering (MAASE '08), ISBN: 978-960-6766-65-7

8. Lukaz A Kurgani and Peter Musilek, "**A survey of knowledge discovery and Data Mining Models**", *The Knowledge Engineering Review,* Vol. 21:1, 1–24. © 2006, Cambridge University Press, doi:10.1017/S0269888906000737

9. Byoung Jik Lee, "**Missing Data Imputation based on Unsupervised Simple Competitive Learning**", Recent Advances in Artificial Intelligence, Knowledge Engineering and Data Bases, ISBN: 978-960-474-154-0

10. Pilar Rey-del-Castillo, and Jesús Cardeñosa, "**Categorical missing data imputation for neural networks with categorical and numeric inputs**", World Academy of Science, Engineering and Technology 55 2009

11. Xiaohua Hu, "**DB-HReduction: A Data Preprocessing Algorithm for data mining applications**", 0893-9659/03

12. Frida Coaquira and Edgar Acuña, "**Applications of Rough Sets Theory in Data Preprocessing for Knowledge Discovery**", Proceedings of the World Congress on Engineering and Computer Science 2007, WCECS 2007, October 24-26, 2007, San Francisco, USA, ISBN:978-988-98671-6-4

13. Helyane Bronoski Borges, and Júlio Cesar Nievola, "**Attribute Selection Methods Comparison for Classification of Diffuse Large B-Cell Lymphoma**", World Academy of Science, Engineering and Technology 8 2005

14. Marco Richeldi and Pier Luca Lanzii, "**Performing Effective Feature Selection by Investigating the Deep Structure of the Data**", From: KDD-96 Proceedings. Copyright © 1996, AAAI (**www.aaai.org**)

15. Asha Gowda Karegowda1, A. S. Manjunath2 & M.A.Jayaram3, "**Comparitive Study of Attribute Selection Using Gain Ratio and Correlation Bases Feature Selection"**, International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No. 2, pp. 271-277

16. Ellen Pitt and Richi Nayak, "**The Use of Various Data Mining and Feature Selection Methods in the Analysis of a Population Survey Dataset**", Second Workshop on Integrating AI and Data Mining (AIDM 2007) Copyright © 2007, Australian Computer Society, Inc

17. P. Ranjit Jeba Thangaiah, R. Shriram, and K. Vivekanandan, "**Adaptive hybrid methods for Feature selection based on Adaptive feature selection**" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009

18. S. B. Kotsiantis, "**Supervised Machine Learning: A Review of Classification Techniques**", Informatica 31 (2007) 249-268

19. Thair Nu Phyu, "**Survey of Classification Techniques in Data Mining**", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong

20. Indranil Bose and Xi Chen, "**Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn**", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009, Vol I, IMECS 2009, March 18 - 20, 2009, Hong Kong

21. Madjid Tavanaa, Prafulla Joglekara, Michael A. Redmondb, "**An automated entity–relationship clustering algorithm for conceptual database design**" doi:10.1016/j.is.2006.07.001, 0306-4379 2006 Elsevier

22. M. Dash 1, H. Liu2, "**Feature Selection for Classification**", Intelligent Data Analysis 1 (1997) 131–156, 1088-467X/97/$17.00 Ó 1997 Elsevier Science

23. Mark A. Hall, Geoffrey Holmes, "**Benchmarking Attribute Selection Techniques for Discrete Class Data Mining**", IEEE Transactions on Knowledge and Data Engineering, VOL. 15, NO. 3, MAY/JUNE 2003

24. Iiawei Han and Michiline Kamber, **Data Mining Concepts and Techniques**, Morgan Kaufmann Publishers, Second Edition, 2006

25. Aswathappa, **Human Resource Management**, Tata McGraw Hill, 3rd Edition, 2002

26. www.ProjectManagementDocs.com

27. www.mathworks.com

28. www.en.wikipedia.org/wiki/Correlation_and_dependence

29. www.statsoft.com/textbook/basic-statistics