SEMINAR PAPER

ON

DATA MINING TECHNIQUES

BY

ESEKU JOAN IBUKUN

MATRIC NO: HCSF/15/0023

SUBMITTED TO SCHOOL OF PURE AND APPLIED SCIENCE

OGUN STATE INSTITUTE OF TECHNOLOGY, IGBESA.

IN PARTIAL FULFILMENT FOR THE AWARD OF HIGHER

NATIONAL DIPLOMA IN THE DEPARTMENT OF COMPUTER SCIENCE.

MAY, 2017

CERTIFICATION

This is to certify that this research work has been carried out by **ESEKU JOAN IBUKUN** with the matric number HCSF/15/0023 in computer science department, **OGUN STATE INSTITUTE OF TECHNOLOGY, IGBESA OGUN STATE.**

…… ……

Eseku Joan Ibukun Date

Student

…… ……

Mr Aluko Date

Supervisor Signature

…… …….

Mrs. Oladejo. R Date

Head of Department Signature

DEDICATION

This project work is dedicated to Almighty God. Also to my parent Mr & Mrs ESEKU for giving me a formal education at all cost.

ACKNOWLEDGEMENT

My utmost gratitude, thanksgiving and appreciation go to the almighty God; I also give thanks to my parent Mr. & Mrs. ESEKU for their moral and financial support for this study to come to completion. I would like to thank my supervisor Mr ALUKO and the HOD computer science department Mrs OLADEJO RACHEAL for giving me such a wonderful opportunity to expand my knowledge for my own branch and giving me guidelines to present a seminar report. It helped me a lot to realize of what we study for.

Secondly, I would like to thank my friends and course mates who helped me as i went through my seminar work and helped to modify and eliminate some of the irrelevant and unnecessary stuffs.

TABLE OF CONTENTS

ABSTRACT

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

Generally data mining contains several algorithms and techniques for picking out interesting patterns from large data sets. Data mining techniques are classified into two categories: supervised learning and unsupervised learning. In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classification, Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining etc. are common examples of supervised learning. In unsupervised learning, we do not create a model or hypothesis prior to the analysis. We just apply the algorithm directly to the dataset and observe the results.

Then a model can be created on the basis of the obtained results. Clustering is one of the examples of unsupervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbors have been used for knowledge discovery from large data sets

CHAPTER ONE

1.0 INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

Data mining can be used to uncover patterns in data but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. Data mining cannot discover patterns that may be present in the larger body of data if those patterns are not present in the sample being "mined". Inability to find patterns may become a cause for some disputes between customers and service providers. Therefore data mining is not foolproof but may be useful if sufficiently representative data samples are collected. The discovery of a particular pattern in a particular set of data does not necessarily mean that a pattern is found elsewhere in the larger data from which that sample was drawn. An important part of the process is the verification and validation of patterns on other samples of data. With the advancement of information technology and increased application of high speed computers in diversified fields, there has been an unexpected growth in the amount of data that is being generated in the repositories. The data may be in the form of audio, text or video also in different format, which is both complex and large. There is a need to convert this data into information by analyzing it from various dimensions, identifying the relationships or grouping the data into categories. This information can then be used in different decision making processes. Data mining is to discover knowledge that is of interest from large amounts of data stored in data repositories. Mining refers to extraction of something which is of value ,hence , data mining is to pull out valuable information from data. Numerous data mining tools are available in the market to predict future trends and assist decision making, that further help organizations to make proactive decisions by looking into past and present data. The varied application areas of data mining are marketing/sales, customer relationship management, banking, insurance, fraud detection, bioinformatics and many more.

The field of data mining is an emerging research area with important applications in Engineering, Science, Medicine, Business and Education. The size of data base in educational application is large where the number of records in a data set can vary from some thousand to thousand of millions. The size of data is accumulated from different fields exponentially increasing. Data mining has been used different methods at the intersection of Machine Learning, Artificial Intelligence, Statistics and Database Systems. The overall aim of the data mining process is to extract information from huge datasets and transform it into understandable structure for further use. Data mining techniques which extract information from huge amount of data have been becoming popular in education domains.

1.1 USES OF DATA MINING

1. **Spatial data mining**: Spatial data mining refers to discovering and extract implicit, significant and unforeseen patterns from large spatial database
2. **Temporal data mining**: The Process of extracting useful and previously undiscovered information from a large dataset, along with some temporal reasoning.
3. **Sequence data mining**: A type of data mining where various patterns which are present in the data are discovered. It could be an association between variables or the way in which events are occurring
4. **Intention mining**: To determine computer user intention from the past data pertaining in interacting with the system

1.2 THE SCOPE OF DATA MINING

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

1. **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

2. **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include

detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

1.3 APPLICATIONS OF DATA MINING

Data mining is a process that analyzes a large amount of data to find new and hidden information that improves business efficiency. Various industries have been adopt data mining to their mission-critical business processes to gain competitive advantages and help business grows. This tutorial illustrates some data mining applications in sale/marketing, banking/finance, health care and insurance, transportation and medicine.

Data Mining Applications in Sales/Marketing

1. Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in prompt and cost effective way. The following illustrates several data mining applications in sale and marketing.
2. Data mining is used for market basket analysis to provide information on what product combinations were purchased together, when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products that they may have been missed or overlooked.
3. Retail companies uses data mining to identify customer’s behavior buying patterns.

Data Mining Applications in Banking / Finance

1. Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection.
2. Data mining is used to identify customers loyalty by analyzing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. The higher of the score, the more relative loyal the customer is.
3. To help bank to retain credit card customers, data mining is applied. By analyzing the past data, data mining can help banks predict customers that likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers.
4. Credit card spending by customer groups can be identified by using data mining.
5. The hidden correlation's between different financial indicators can be discovered by using data mining.
6. From historical market data, data mining enables to identify stock trading rules.

Data Mining Applications in Health Care and Insurance

The growth of the insurance industry entirely depends on the ability of converting data into the knowledge, information or intelligence about customers, competitors and its markets. Data mining is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully. The data mining applications in insurance industry are listed below:

1. Data mining is applied in claims analysis such as identifying which medical procedures are claimed together.
2. Data mining enables to forecasts which customers will potentially purchase new policies.
3. Data mining allows insurance companies to detect risky customers' behavior patterns.
4. Data mining helps detect fraudulent behavior.

Data Mining Applications in Transportation

1. Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

Data Mining Applications in Medicine

1. Data mining enables to characterize patient activities to see incoming office visits.
2. Data mining helps identify the patterns of successful medical therapies for different illnesses.
3. Data mining applications are continuously developing in various industries to provide more hidden knowledge that increases business efficiency and grows businesses.

1.4 ADVANTAGES OF DATA MINING

Data mining is an important part of knowledge discovery process that we can analyze an enormous set of data and get hidden and useful knowledge. Data mining is applied effectively not only in business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government…etc. Data mining has a lot of advantages when using in a specific industry. Besides those advantages, data mining also has its own disadvantages e.g., privacy, security and misuse of information. We will examine those

A. Marketing / Retail

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers. Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

B. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

C. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipment's and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

D. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities. www.studymafia.org

1.5 DISADVANTAGES OF DATA MINING

A. Privacy Issues

The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time the personal information they own probably is sold to other or leak.

B. Security issues

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

C. Misuse of information/inaccurate information

Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people

CHAPTER TWO

2.0 LITERATURE REVIEW

Han et al (2001). provided a comprehensive survey, in database perspective, on the data mining techniques developed recently. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining, was reviewed by them. Techniques for mining knowledge in different kinds of databases, included relational, transaction, object oriented, spatial, and active databases, as well as global information systems, was examined by them. Clustering is the most commonly used technique of data mining under which patterns are discovered in the underlying data. Sidhu et al (2001). presented that how clustering was carried out and the applications of clustering. They also provided us with a framework for the mixed attributes clustering problem and also showed us that how the customer data can be clustered identifying the high-profit, high-value and low-risk customer.

Huang presented an algorithm, called k-modes, to extend the k-means paradigm to categorical domains. He introduced new dissimilarity measures to deal with categorical objects, replace means of clusters with modes, and use a frequency based method to update modes in the clustering process to minimize the clustering cost function. Experimented on a very large health insurance data set consisted of half a million records and 34 categorical attributes showed that the algorithm was scalable in terms of both the number of clusters and the number of records Berkhin (1990) surveyed that concentrated on clustering algorithms from a data mining perspective. K-prototypes algorithm was proposed, which was based on the k-means paradigm but removed the numeric data limitation whilst preserved its efficiency. In that algorithm, objects were clustered against k prototypes. A method was developed to dynamically update the k prototypes in order to maximize the intra cluster similarity of objects.

The efficiency and scalability issues were addressed by proposing a data classification method which integrated attribute oriented induction, relevance analysis, and the induction of decision trees. Such an integration lead to efficient, high quality, multiple level classification of large amounts of data, the relaxation of the requirement of perfect training sets, and the elegant handling of continuous and noisy data.

Antonie et al. presented some experiments for tumor detection in digital mammography. They investigated the use of different data mining techniques, for anomaly detection and classification. The experiments they conducted demonstrated the use and effectiveness of association rule mining in image categorization. Kohavi et al. described the two most commonly used systems for induction of decision trees for classification: C4.5 and CART. They highlighted the methods and different decisions made in each system with respect to splitting criteria, pruning, noise handling, and other differentiating features.

Ma et al. proposed to integrate classification and association rule mining techniques. The integration was done by focusing on mining a special subset of association rules, called class association rules (CARs). An efficient algorithm was also given to build a classifier based on the set of discovered CAR. In two new algorithms were presented for solving the problem of discovering association rules between items in a large database of sales transactions. The algorithms were fundamentally different from the known algorithms. They showed how the best features of the two proposed algorithms could be combined into a hybrid algorithm, called Apriori Hybrid. Agrawal et al. proposed an efficient algorithm that generated all significant association rules between items in the database. The algorithm incorporated buffer management and novel estimation and pruning techniques. Yavas et al. proposed an algorithm for predicting the next inter-cell movement of a mobile user in a Personal Communication Systems network. The performance of the proposed algorithm was evaluated through simulation as compared to two other prediction methods.

Nanopoulos et al. presented a new context for the interpretation of Web pre fetching algorithms as Markov predictors. They identify the factors that affect the performance of Web pre fetching algorithms and proposed a new algorithm called WM, which was based on data mining and was proved to be a generalization of existing ones. Velmurgun T. et al. attempted to analyze performance of K-means and Fuzzy C-means clustering techniques in the field of data mining. The performance compared on the basis of clustering result quality. Kavitha P., T. Sasipraba (2001) evaluated the performance of distributed data mining framework on Java platform. Association rule mining was used for discovering interesting patterns from a large amount of data. Yujie Zheng proposed a methodology for clustering in data mining to improve the standard of higher education used to find data segmentation and pattern information. M.Sukanya et al. used classification and clustering algorithms of data mining for the performance improvement in

education sector. By using these algorithms an educational institute could predict the number of enrolled students. Manoj Bala et al. applied an application of data mining in educational institute to extract the useful information from the huge dataset and provided analytical tool to view and used this information for decision making process. They also conducted a research on student learning result based on data mining. Hamidah J (2003) used a potential classification technique for academic talent forecasting in higher educational institutes. He proposed a classification model to increase the academic talent in higher educational institutions.

Tai Chang (1996) applied data mining techniques to analyze the course preferences and course completion rates of enrollees in extension education courses at a University in Taiwan. Some of the data mining algorithms like decision tree, link analysis, and decision forest were used for further analysis.

2.1 KNOWLEDGE DISCOVERY IN DATABASES

Knowledge discovery in databases (KDD) can be defined as a process of obtaining information using data registered in a databank, an implicit, previously unknown, potentially useful and understandable knowledge. The expression Data Mining (DM) first appears, as a synonym of KDD, but it is only one of the stages of the knowledge discovery in databases in the KDD global process. The knowledge that is possible to acquire through the DM has been very useful in the most different areas, such as medicine, finances, commerce, marketing, telecommunications, meteorology, agriculture and cattle rising, bioinformatics, among others. Data mining is not a trivial process; it consists of the ability to identify, in the data, the valid, new, potentially useful and understandable patterns, involving statistical methods, visualization tools and artificial intelligence techniques So, the KDD process uses databases concepts, statistical methods, visualization tools and artificial intelligence techniques, dividing into the following phases: selection, pre-processing, transformation, DM and evaluation/interpretation. Among these phases, the most important is the data mining, focuses of countless studies in several areas of knowledge that confirms the assumption of the changing of data into information and later into knowledge, which makes the technique necessary for the making-decision process.

Data mining has several phases: the clear definition of the problem; the selection of all the internal and external data sources and the preparation of data, which includes the pre-processing, data reformatation and analysis of the results taken from the DM process.

CHAPTER THREE
METHODOLOGY/ARCHITECTURE

3.0 DATA MINING TECHNIQUES

Generally data mining contains several algorithms and techniques for picking out interesting patterns from large data sets. Data mining techniques are classified into two categories: supervised learning and unsupervised learning.

In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classification, Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining etc. are common examples of supervised learning. In unsupervised learning, we do not create a model or hypothesis prior to the analysis. We just apply the algorithm directly to the dataset and observe the results. Then a model can be created on the basis of the obtained results. Clustering is one of the examples of unsupervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbors have been used for knowledge discovery from large data sets. Some of the common and useful data mining techniques are:

1. CLASSIFICATION

Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level. Classification algorithm requires that the classes be defined based on the data attribute value. It describes these classes according to the characteristics of the data that is already known to belong to the classes. The classifier training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination. Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups.

For example, we can apply classification in application that “given all records of employees who left the company, predict who will probably leave the company in a future period.” In this case, we divide the records of employees into two groups that named “leave” and “stay”. And then we can ask our data mining software to classify the employees into separate groups.

Classification is learning rules that can be applied to new data and will typically include following steps: preprocessing of data, designing modelling, learning/feature selection and validation /evaluation. Classification predicts categorical continuous valued functions. For example, we can make classification model to categorize bank loan application as either safe or risky. Classification is the derivation of model which determines the class of an object based on its attributes. A set of object is given as training set in which every object is represented by vector of attributes along with its class. By analysing the relationship between attributes and class of the objects in the training set, classification model can be constructed. Such classification model can be used to classify future objects and develop a better understanding of the classes of the objects in the data base.

Types of classification Techniques:

1. Classification by decision tree induction
2. Bayesian Classification
3. Neural Networks
4. Support Vector Machines (SVM)
5. Classification Based on Associations

2. CLUSTERING

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

Clustering is identifying similar groups from unstructured data. Clustering is the task of grouping a set of objects in a such a way that object in same group are more similar to each other than to

those in other groups. Once the clusters are decided, the objects are labelled their corresponding clusters, and common features of the objects in cluster are summarized to form a class description. For example, a bank may cluster its customer in to several groups based on the similarities of their income, age, sex, residence etc, and the command characteristics of the customers in a group can be used to describe that group of customers.

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in library as an example. In a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for entire library.

Types of clustering methods

1. Partitioning Methods
2. Hierarchical Agglomerative (divisive) methods
3. Density based methods
4. Grid-based methods

3. DECISION TREE

A decision tree is a flow chart like tree structure, where each node denotes test on an attribute value, each branch represents the result of the test, and tree leaves represent classes. The drive model can be represented in different forms such as classification (If-Then) rules, decision tree, mathematical formula or neural networks. Decision tree can easily be converted to classification tree. Decision trees are simple to understand and provide good results even with small data. Decision tree induction algorithms can be used for classification in many application areas, such as Education, Medicine, Manufacturing, Production, Financial analysis, Fraud Detection and Astronomy etc. There are several data mining algorithms such as C4.5, ID3, CART, J48, NB Tree, REP Tree etc.

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers.

Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

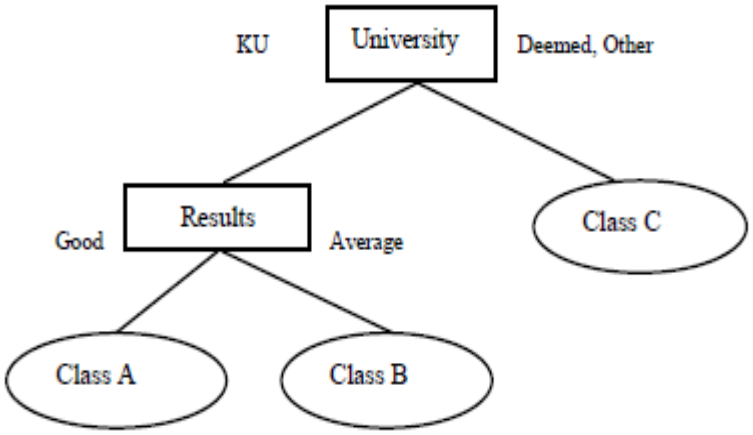

Fig 3.1 Decision Tree

4. PREDICTION

Prediction is a data mining technique that is used to identify the relationship between independent variables and relationship between dependent and independent variables. Prediction analysis can be used in education domains. Regression technique can be used to generate a model for prediction. Regression analysis can be used to model the relationship between one or more independent variable and dependent variables. Prediction techniques can be used to predict the possible values of some missing data and the value distribution of certain attributes in a set of objects. It finds the attribute related to the interest and predicting the value distribution based on the set of data similar to the selected objects. There are many regression techniques such as Linear Regression, Nonlinear Regression, Multivariate Linear Regression, and Multivariate Nonlinear Regression. The prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables.

For instance, the prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based

on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

5. ASSOCIATION RULE

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and therefore they can put beers and crisps next to each other to save time for customer and increase sales. Association is looking for relationship between variables or objects. It aims to extract interesting association, correlations or casual structures among the objects i.e. the appearance of another set of objects in. The association rules can be useful for marketing, commodity management, advertising etc. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.

It is intended to identify strong rules discovered in databases using different measures of Interestingness and based on the concept of strong rules presented in, introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets.

Types of association rule

1. Multilevel association rule
2. Multidimensional association rule
3. Quantitative association rule

6. SEQUENTIAL PATTERNS

It is one of the data mining techniques that seek to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data. Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

7. NEURAL NETWORKS

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

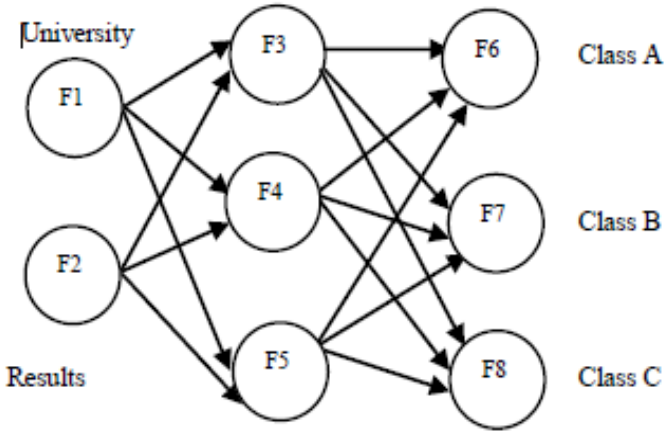

Fig 7.1 Neural Networks

The area of neural networks probably belongs to the border line between the artificial intelligence and approximation algorithm. A neural network is a collection of neurons like processing units with weighted connection between the units. It composes of many elements, called nodes which are connected in between. The connection between two nodes is weighted and by the adjustment of this weight, the training of the network is performed. A classification model can be represented in different forms like neural network and decision tree which is shown in fig. 3.1 and fig.7.1. There are many advantages of neural networks such as adaptive learning ability, self-organization, real time operation and insensitivity to noise. Neural networks are used to identifying patterns or trends in data and well suited for prediction or forecasting needs. There are several neural network algorithms such as Back Propagation, NN Supervised Learning, and Radial Base Function (RBF) Network etc.

CHAPTER FOUR
SUMMARY AND CONCLUSION

4.0 SUMMARY

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides.

Generally data mining contains several algorithms and techniques for picking out interesting patterns from large data sets. Data mining techniques are classified into two categories: supervised learning and unsupervised learning. In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classification, Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining etc. are common examples of supervised learning. In unsupervised learning, we do not create a model or hypothesis prior to the analysis.

4.1 CONCLUSION

This seminar work briefly reviewed data mining techniques. Proper selection of data mining technique and domain knowledge is an important consideration to make effective utilization of data mining. This work will provide a pathway for beginners in this area. In this seminar, we have discussed detail study of data mining with various studies like tasks, tools and techniques. The implementation of data mining techniques will allow users to retrieve meaningful information from virtually integrated data. These techniques provide variety of applications for industries like retail, telecommunication, Bio-medical etc. These tools predict future trends and

behaviors, allowing business to make proactive and present knowledge in the form which is easily understood to human.

REFERENCES

Bharati M. (2001), "Data Mining Techniques and Application", Indian Journal of Computer Science and Engineering, Vol. 1 No. 4; pp. 301-305.

D.W. Cheung, S.D.Lee and B.Kao (1992), "A general incremental technique for maintaining discovered association rule". Proc. In fifth international conference on data base system for advanced applications, Australia, 1997.

G.Piatetsky-shapiro, U.Fayyed and P.Smith.(2001, June 3) From data mining to Knowledge discovery: An overview. Advances in knowledge Discovery and Data Mining, pages 1-35, MIT Press, 1996.

Kavitha P.and T. Sasipraba (1998), "Performance Evaluation of Algorithms using a Distributed Data Mining Framework based on Association Rule Mining", International Journal on Computer Science & Engineering (IJCSE), 2011.

M.Sukan et al (2000), "Data Mining: Performance Improvement in Education Sector using Classification and Clustering Algorithm", International Conference of Computing and Control Engineering (ICCCE) 12-13 April, 2012.

Velmurugan T. et al (2010), "Performance Evaluation of K-Means & fuzzy C-means Clustering Algorithm for Statistical Distribution of Input Data Points", European Journal of Scientific Research, Vol. 46, 2010.

Yujie Zheng F (1997,March 15) "Clustering Methods in Data Mining with its Application in Higher Education", International Conference on Education Technology and Computer, Vol. 43, 2012, IACSIT Press, Singapore.