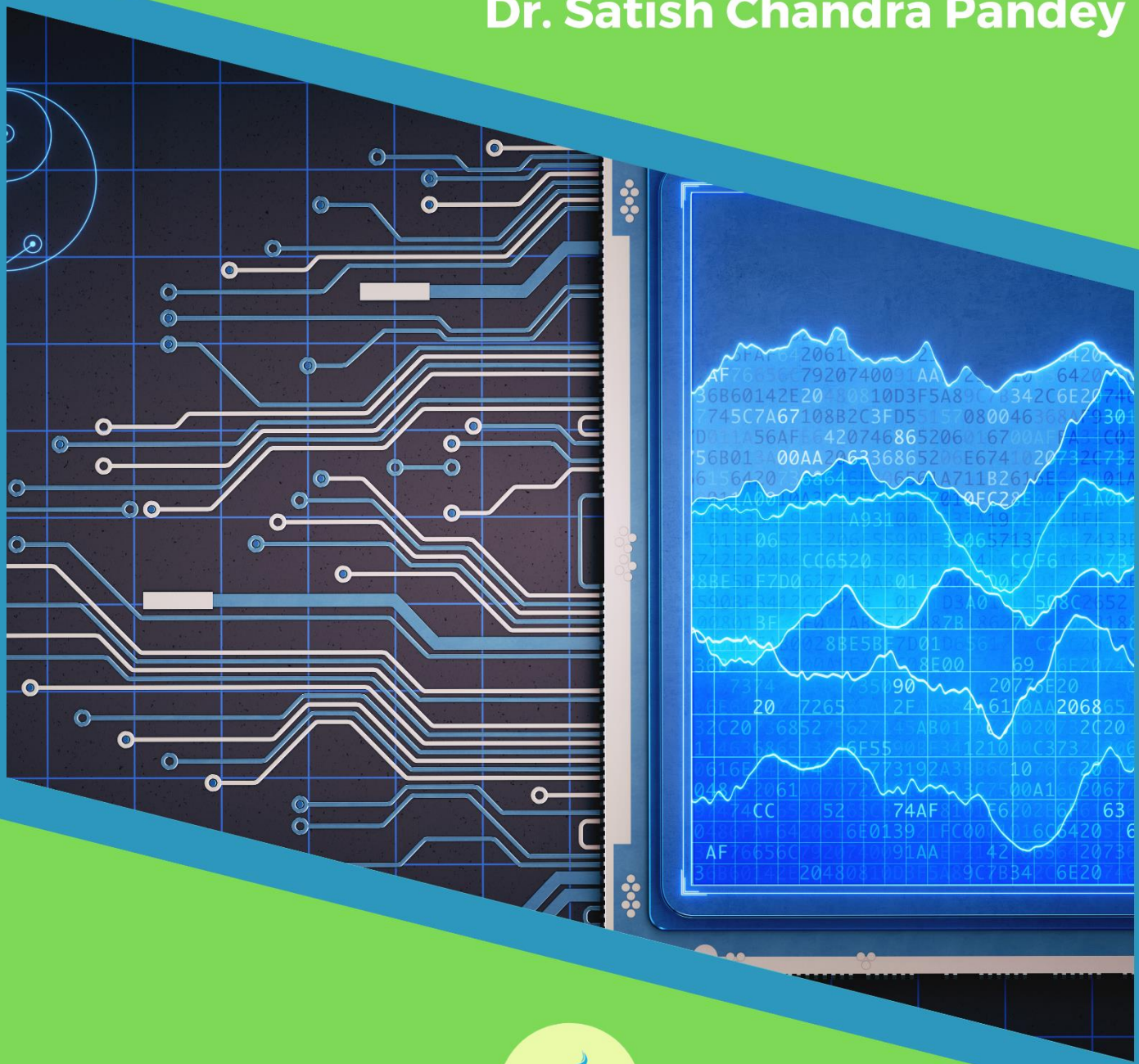
ISBN: 978-93-88901-29-1

DATA MINING TECHNIQUES & APPLICATIONS

Dr. Satish Chandra Pandey



Bhumi Publishing

First Edition: 2022

Data Mining Techniques & Applications

(ISBN: 978-93-88901-29-1)

Dr. Satish Chandra Pandey

Bhumi Publishing

2022

First Edition: December, 2022

ISBN: 978-93-88901-29-1

Published by:

Bhumi Publishing,

Nigave Khalasa, Kolhapur 416207, Maharashtra, India

Website: www.bhumipublishing.com

E-mail: bhumipublishing@gmail.com

Book Available online at:

https://www.bhumipublishing.com/book/

PREFACE

My first experience with Data Mining Techniques & Applications. It is the ultimate challenge for a learner, encompassing everything from low level device manipulation, to concurrency to object oriented design. This book is intended as a text in Data Mining Techniques & Applications for Engineering & Post Graduate Level Students. We have attempted to cover the major topics in Data Mining Techniques & its Applications in depth.

Coverage

A brief synopsis of the chapters and comments on their appropriateness in a basic course in Therefore appropriateness in basic course is therefore appropriate.

In Chapter-I General Introduction, various definition of Data mining and details of literature review are described. In Chapter-II Data Processing & Cleaning, Data mining process & its Types are described. In Chapter-III Data Integration, Data Integration Approaches, Issues in Data Integration & Techniques of Data Integration are described. In Chapter-IV Data Reduction, Techniques of Data Reduction are described. In Chapter-V Methods & Methodology, steps of Data mining, Data mining process models and various Data mining techniques are described with corresponding Figures. In Chapter-VI Data Mining Techniques, Data mining technique has been applied in healthcare domain to find out how certain variables are associated with the onset of diabetics. Data mining technique have been applied in the marketing domain to direct marketing strategy because due to competitive marketing environment advancement in technology and changing behavior of customers which are difficult to predict. Various Data mining techniques have been applied in the field of terror related activities. In Chapter-VII Applications of Data Mining, Data mining Application in various domains and limitations of Data mining are pointed out.

Our aim is to present these concepts and algorithms in a general a large number of examples that pertain to the most popular and the most innovative Data Mining. This book explores the Data Mining Techniques & Applications. It is intended for those wishing to learn more about Data Mining in general or for those with interest in a particular system who desire a broader perspective on its operation.

I wish to thank Managing Team & all the staff at Bhumi Publishing, Kolhapur, Maharashtra, India who helped produce this effort.

The book concentrates on generally applicable design features and does not cover more specialized topics. Although it is hoped all material in this book is accurate, the possibility does exist that some errors are present. Notification of errors, omissions or suggested improvements can be sent to pandey.satishchandra@gmail.com.

- Dr. Satish Chandra Pandey

ACKNOWLEDGEMENT

I first of all record our Gratitude to the Authorities & Dean Academics of Jayoti Vidyapeeth Women's University Jaipur whose efforts maintained the progress of my work. I thank my all students and my colleagues at Jayoti Vidyapeeth Women's University Jaipur for their understanding and support.

I am especially indebted to my mother Late Smt. Raj Kumari Pandey for her continuous inspiration, encouragement and best blessings.

I express my sincere regards to My Brother in Law Late Dr. Diwaker Tiwari, Scientist, IIG, Navi Mumbai, for their affection, constant encouragement and moral support during my all Academic Work.

I also express my deepest sense of feelings to My Brother in Law Late Shri Awadhesh Kumar Pandey, Assistant Registrar IGNOU, Varanasi and Dr. Gayatri Pandey, My Elder sister for their steamed cooperation, & encouragement support from time to time during the period of my work.

I am specially indebted to my elder sister Dr. Nirupama Tiwari, Scientific Officer Grade- F, BARC, Mumbai for her affection & academic suggestions without which the present investigations could not have been completed.

I also express my heartiest thanks to my wife Smt Sweta Pandey, Sr.Teacher, Children Academy, Alwar (Raj.) & My son Rudra Pratap Pandey for emotional support & painstaking cooperation with their critical review was invaluable.

Sincere thanks to my Father in law Shri Onkar Nath Pandey & Mother in law Smt Sheela Pandey for their blessings, encouragement and valuable help during these tenure.

I express my heartiest thanks to Dr. Y. Mishra Sir, Ex-HOD Physics.Deptt.& Shri K.S.Pandey Sir,Asstt.Prof.CS Deptt,S.H.Kishan PG.College ,Basti (UP) for their blessings.

I wish to thank my colleagues, and friends, for their academic support and valuable suggestions during the period of writing work.

I wish to thank my all elder brothers Ram Chandra Pandey, Satya Prakash Pandey, Dr. J.P.Pandey,Surya Prakash Pandey, Ajay Kumar Pandey, Dr. R.K.Pandey, and other family members whose affections, blessings and supports enabled me to complete this work successfully.

Last but not the least, how to express my sentiments for my father Dr. R.C Pandey, Rtd Principal,S.H K. PG College,Basti, , for help in the collection of Writing materials & documentations, Academic suggestions , financial help and motivation.

Once again we acknowledge our all family members and wish them to know that my thanks is not merely "proforma", they are sincerely offered and they well deserved.

CONTENT

Dedicated to

My Beloved Parents

Late Smt. Raj Kumari Pandey

&

Dr. R. C. Pandey

CHAPTER – 1
GENERAL INTRODUCTION

1.1 Introduction

Industrialization is an important aspect of human civilization and is most essential for the development of any country. These industries generate a huge amount of data every day.

Besides these, marketing sectors, hospitals and clinics, banking & insurance sectors, educational institutions, scientific labs, daily communication and electronic areas, agriculture and forestry, Research & development organizations and government sectors and sectors of terror related activities all produce extremely large datasets per day. The amount of data generated is found to be almost doubled every nine months. Within these mounds of data lies hidden and useful information and knowledge of strategic importance which are extremely difficult to achieve by traditional methods. Manual data analysis has been around for some time now but is creates a bottle neck for large data analysis. Some initial knowledge is known about the data but Data mining could help in a more in-depth knowledge about the data. Data mining as we use the term is a powerful technique to explore and analyze such a large amount of data in order to discover meaningful patterns and rules.

Thus the major reason why Data mining has attracted a great deal of attention in the information industry and in society as whole in recent years is due to the wide availability of huge amounts of data and the eminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection and customer retention to production control and science exploration.

Moreover, for developing computer science and engineering techniques & methodology new demands are generated. Data mining techniques are now being applied to all kinds of domains which are rich in data e.g. image mining and gene data analysis.

1.2 Data mining Definitions

Some of the numerous definitions of data mining or knowledge discovery in data bases are mentioned below:

Data mining also known as knowledge discovery in databases (KDD) is the automated extraction of novel, understandable and potentially useful patterns implicitly stored in large

databases, data warehouses and other massive information repositories comprising textual, numerical, graphical and spatial data.

Data mining or knowledge discovery in databases (KDD) as it is also known is the nontrivial extraction of implicit, previously unknown and potentially useful information from data. This encompasses number of different technical approaches such as clustering, data summarization, and learning classification rules finding dependency networks, analyzing changes and detecting anomalies."

According to Fayyed *et al.* (1996) "Data mining is a non trivial process of discovering novel, implicit, useful and comprehensive knowledge from a large amount of data."

Data mining has also been defined as the process of finding previously unknown patterns and trends in databases & using that information to build predictive models (Kincade, 1998).

Milley (2000) has defined Data mining as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns.

Data mining as the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of statistics, machine learning and database management system.

Data mining according to Kreuze (2001) aim to identify valid, novel, potentially useful, and understandable correlations and patterns in data by combining through copious data sets to sniff out patterns that are too subtle or complex for humans to detect.

Data mining is the search for relationships and global patterns that exist in large databases but are "hidden" among the vast amount of data, such as a relationship between patient data and their medical diagnosis.

Data mining according to Reza Fadai (2002) deals with the discovery of unexpected pattern and new rules that are "hidden" in large databases. It serves as an automated tool that uses multiple advanced computational techniques including artificial intelligence.(The use of computers to perform logical functions) to fully explore and characterize large data sets involving one or more data sources identifying significant recognizable patterns, trends and relationship not easily detected through traditional analytical techniques alone. Data mining has also been defined as the process of discovering meaningful new correlation patterns and trends by sniffing through large amount of data stored in repositories using pattern reorganization techniques as well as statistical and mathematical techniques (Pujari, 2008).

Data mining is the process of extracting previously unknown valid & actionable information from large data (as transaction data or database or data warehouse) & then using the information so derived to make critical business & strategic decisions.

Data mining describes a collection of techniques that aim to find useful but undiscovered pattern in collected data. It‘s goal is to create modules for decision making that predict future behavior based on analysis of past activity. It refers to the process of systemically analyzing large databases to find useful patterns.

1.3 Roots of Data Mining

Roots of Data mining can be tracking back along three lines.

1.3.1 Statistics

The most important line is statistics without statistics there would be no Data mining because statistics are the foundation of most technologies on which Data mining is constructed. Statistics embrace concepts like regression analysis standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis and confidence intervals all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analysis are underpinned. Undoubtedly classical statistical analysis plays an important role in today’s Data mining tools and techniques.

1.3.2 Artificial Intelligence & Machine learning

The second longest family line of Data mining is artificial intelligence and machine learning.AI (Advance Information) is constructed upon heuristics as opposed to statistics and tries to apply human thought like processing to statistical problems. Since this approach requires vast computer processing power, it was not practical until the early 180 when computers began to after useful power at nominal prices.AI obtained a few applications at the very high and scientific government markets but the required supercomputers of the time priced AI out of reach of common people. Machine learning could be considered as an evolution of AI since it blends AI heuristics with advanced statistical methods.

1.3.3 Databases

Data mining’s third family line is databases. The vast amount of data required to be deposited in a repository which needs to be managed. Earlier data was managed in records and fields. Then in several models such as hierarchical network etc. Relational model served the needs of data storage for long time while other advanced system which emerged is object relational databases. But volume of data is very high in Data mining hence we need specialized

servers for it. We call the term as data ware housing. Data warehousing also supports OLAP operations to be applied on it to support decision making (Chen 1999).

1.3.4 Other technologies

Apart from above Data mining includes various other areas e.g. pattern discovery, visualization, business intelligence etc.

1.4 Types of Data to be mined

Data mining is not specific to one type of data but can be applicable to any kind of information repository. However algorithms and approaches may differ when applied to different types of data. Data mining is being put into use and it did for following types of databases.

1.4.1 Flat Files

Flat files are simple data files in next or binary format with a structure known by the Data mining algorithm to be applied.The data in these files can be transactions, time series data scientific measurement etc.

1.4.2 Relational Databases

A relational database consists of a set of tables which have column and rows, where columns represent attributes and rows represent tuples.

1.4.3 Data warehouse

A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified scheme. A data warehouse is helpful in analyzing the data from different sources under the same roof.

1.4.4 Transaction Database

A transaction database is a set of records representing functions, each with a time setup an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. Since relational databases do not allow nested tables (i.e. a set as attribute value), transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and one for the truncation items.

1.4.5 Multimedia Databases

Multimedia databases include video, images, audio and text media. They can be stored on extended object relational or object oriented databases or simply on a file system. Multimedia is characterized by its high dimensionally which makes Data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation and natural languages processing methodologies.

1.4.6 Spatial Databases

Spatial Databases are databases that in addition to usual; data store geographical information like maps, and global or regional positioning.

1.4.7 Time Series Databases

Time series databases contain time related data such as stock market data or logged activities. These databases usually have a continuous flow of new data coming in which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes of different variables as well as the prediction of trends and movements of the variables in time.

1.4.8 World Wide Web

Data in the www is organized in inter connected documents, these documents can be text, audio, video, raw data and even applications conceptually, the www is comprised of three major components. The content of the web which encompasses documents available; the structure of the web, which converts the hyperlinks and the relationship between documents and the usage of the web, describing how and when the resources are accessed. Data mining in the www or web mining tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining. The www is the most important data collection regularly used for reference because of the broad Varity of topics covered and the infinite contributions of resources and publishers.

1.5 Sources of Information

The various sources of information collected in digital form in databases and in flat files are mentioned below:

1.5.1 Business Transactions

Transactions in the business industry are usually time related and can be inter business deal such as purchases exchanges, banking stock etc or intra business operations such as management in house wares and assets.

1.5.2 Scientific Data

Our society is amassing colossal amounts of scientific data that need to be analyzed. Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated.

1.5.3 Medical & Personal Data

Government companies and organizations such as hospitals are stock billing very important quantities of personal data to help them manage human resources better understand a market or simply assist clients. Regardless of the privacy issues the type of data often reveals, this information is collected used and even shared.

1.5.4 Satellite Sensing

There are a countless number of satellites around the globe some are geostationary above a region and some are orbiting amount the earth but all are sending a nonstop stream of data to the surface. NASA, which controls a large number of satellites, receives more data every second than what all NASA reaches and engineers can cope with.

1.5.5 Games & Digital media

Our society is collecting a tremendous amount of data and statistics about games, plaster and athletes from hockey scores, basketball passes and car racing lapses to swimming times boxers pushes and chess positions all the data are stored.

Many radio stations, televisions channels and film studios are digitizing their audio and video collectors to improve the management of their multimedia assets.

1.5.6 CAD and Software Engineering Data

There are a multitude of computer assisted design (CAD) systems for architects to design buildings or engineers to conceive system components or circuits. These systems are generating a tremendous amount of data. Moreover S/W engineering is a source of considerable similar data with code, function libraries objects etc, which need powerful tools for management and maintenance.

1.5.7 Email messages

Most of the communications within and between companies or research organizations or even private people, are based on reports and memos in textual forms often exchanged by email. These messages are regularly stored in digital form for future use and reference creating formidable digital libraries.

CHAPTER – 2
DATA PROCESSING & CLEANING

2.1 Introduction

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. "How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process?" There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. Data transformations, such as normalization, may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements. These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format. Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining.

In this chapter, we introduce the basic concepts of data preprocessing in following Section. The methods for data preprocessing are organized into the following categories: data cleaning, data integration, data reduction, and data transformation.

Data processing is collecting raw data and translating it into usable information. The raw data is collected, filtered, sorted, processed, analysed, stored, and then presented in a readable format. It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization.

The data processing is carried out automatically or manually. Nowadays, most data is processed automatically with the help of the computer, which is faster and gives accurate results. Thus, data can be converted into different forms. It can be graphic as well as audio ones. It depends on the software used as well as data processing methods.

After that, the data collected is processed and then translated into a desirable form as per requirements, useful for performing tasks. The data is acquired from Excel files, databases, text file data, and unorganized data such as audio clips, images, GPRS, and video clips.

Data processing is crucial for organizations to create better business strategies and increase their competitive edge. By converting the data into a readable format like graphs, charts, and documents, employees throughout the organization can understand and use the data.

2.2 Types of Data Processing

There are different types of data processing based on the source of data and the steps taken by the processing unit to generate an output. There is no one size fits all method that can be used for processing raw data, these are:

2.2.1 Batch Processing

In this type of data processing, data is collected and processed in batches. It is used for large amounts of data. For example, the payroll system.

2.2.2 Single User Programming Processing

It is usually done by a single person for his personal use. This technique is suitable even for small offices.

2.2.3 Multiple Programming Processing

This technique allows simultaneously storing and executing more than one program in the CPU. Data is broken down into frames and processed using two or more CPU’s within a single computer system. It is also known as parallel processing. Further, the multiple programming techniques increase the respective computer's overall working efficiency. A good example of multiple programming processing is weather forecasting.

2.2.4 Real-time & Online Processing

This technique facilitates the user to have direct contact with the computer system. This technique eases data processing. This technique is also known as the direct mode or the interactive mode technique and is developed exclusively to perform one task. It is a sort of online processing, which always remains under execution. For example, withdrawing money from ATM.

Online Processing facilitates the entry and execution of data directly; so, it does not store or accumulate first and then process. The technique is developed to reduce the data entry errors, as it validates data at various points and ensures that only corrected data is entered. This technique is widely used for online applications. For example, barcode scanning.

2.2.5 Time-sharing Processing

This is another form of online data processing that facilitates several users to share the resources of an online computer system. This technique is adopted when results are needed swiftly. Moreover, as the name suggests, this system is time-based. Following are some of the major advantages of time-sharing processing, such as:

- Several users can be served simultaneously.
- All the users have an almost equal amount of processing time.
- There is a possibility of interaction with the running programs.

2.2.6 Distributed Processing

This is a specialized data processing technique in which various computers (located remotely) remain interconnected with a single host computer making a network of computers. All these computer systems remain interconnected with a high-speed communication network. However, the central computer system maintains the master database and monitors accordingly. This facilitates communication between computers.

2.3 Examples of Data Processing

Data processing occurs in our daily lives whether we may be aware of it or not. Here are some real-life examples of data processing, such as:

- Stock trading software that converts millions of stock data into a simple graph.
- An e-commerce company uses the search history of customers to recommend similar products.
- A Digital marketing company uses demographic data of people to strategize location-specific campaigns.
- A Self-driving car uses real-time data from sensors to detect if there are pedestrians and other cars on the road.

2.4 Data Cleaning in Data Mining

Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data Cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is the main issue in quality information management. Data quality problems occur anywhere in information systems. These problems are solved by data cleaning.

Data cleaning is fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. If data is incorrect, outcomes and algorithms are unreliable,

even though they may look correct. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.

Generally, data cleaning reduces errors and improves data quality. Correcting errors in data and eliminating bad records can be a time-consuming and tedious process, but it cannot be ignored. Data mining is a key technique for data cleaning. Data mining is a technique for discovering interesting information in data. Data quality mining is a recent approach applying data mining techniques to identify and recover data quality problems in large databases. Data mining automatically extracts hidden and intrinsic information from the collections of data. Data mining has various techniques that are suitable for data cleaning.

Understanding and correcting the quality of your data is imperative in getting to an accurate final analysis. The data needs to be prepared to discover crucial patterns. Data mining is considered exploratory. Data cleaning in data mining allows the user to discover inaccurate or incomplete data before the business analysis and insights.

In most cases, data cleaning in data mining can be a laborious process and typically requires IT resources to help in the initial step of evaluating your data because data cleaning before data mining is so time-consuming. But without proper data quality, your final analysis will suffer inaccuracy, or you could potentially arrive at the wrong conclusion.

2.5 Steps of Data Cleaning

While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to cleaning your data, such as:

2.5.1 Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze.

For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient, minimize distraction from your primary target, and create a more manageable and performable dataset.

2.5.2 Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. For example, you may find "N/A" and "Not Applicable" in any sheet, but they should be analyzed in the same category.

2.5.3 Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data entry, doing so will help the performance of the data you are working with.

However, sometimes, the appearance of an outlier will prove a theory you are working on. And just because an outlier exists doesn't mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

2.5.4 Handle missing data

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered, such as:

You can drop observations with missing values, but this will drop or lose information, so be careful before removing it.

You can input missing values based on other observations; again, there is an opportunity to lose the integrity of the data because you may be operating from assumptions and not actual observations.

You might alter how the data is used to navigate null values effectively.

2.5.5 Scrub for duplicate data

Determine duplicates to save time when analyzing data. Frequently attempted the same data can be avoided by analyzing and investing in separate data erasing tools that can analyze rough data in quantity and automate the operation.

2.5.6 Validate and QA

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation, such as:

- Does the data make sense?
- Does the data follow the appropriate rules for its field?

- Does it prove or disprove your working theory or bring any insight to light?
- Can you find trends in the data to help you for your next theory?

  If not, is that because of a data quality issue?

Because of incorrect or noisy data, false conclusions can inform poor business strategy and decision-making. False conclusions can lead to an embarrassing moment in a reporting meeting when you realize your data doesn't stand up to study. Before you get there, it is important to create a culture of quality data in your organization. To do this, you should document the tools you might use to create this strategy.

2.6 Process of Data Cleaning

The following steps show the process of data cleaning in data mining.

2.6.1 Monitoring the errors

Keep a note of suitability where the most mistakes arise. It will make it easier to determine and stabilize false or corrupt information. Information is especially necessary while integrating another possible alternative with established management software.

2.6.2 Standardize the mining process

Standardize the point of insertion to assist and reduce the chances of duplicity.

2.6.3 Validate data accuracy

Analyze and invest in data tools to clean the record in real-time. Tools used Artificial Intelligence to better examine for correctness.

2.6.4 Research on data

Before this activity, our data must be standardized, validated, and scrubbed for duplicates. There are many third-party sources, and these Approved & authorized parties sources can capture information directly from our databases. They help us to clean and compile the data to ensure completeness, accuracy, and reliability for business decision-making.

2.6.5 Communicate with the team

Keeping the group in the loop will assist in developing and strengthening the client and sending more targeted data to prospective customers.

CHAPTER – 3
DATA INTEGRATION

3.1 Data Integration in Data Mining

Data integration is the process of merging data from several disparate sources. It has been an integral part of data operations because data can be obtained from several sources. It is a strategy that integrates data from several sources to make it available to users in a single uniform view that shows their status. While performing data integration, you must work on data redundancy, inconsistency, duplicity, etc. In data mining, data integration is a record pre-processing method that includes merging data from a couple of the heterogeneous data sources into coherent data to retain and provide a unified perspective of the data.

There are communication sources between systems that can include multiple databases, data cubes, or flat files. Data fusion merges data from various diverse sources to produce meaningful results. The consolidated findings must exclude inconsistencies, contradictions, redundancies, and inequities.

Data integration is important because it gives a uniform view of scattered data while also maintaining data accuracy. It assists the data-mining program in meaningful mining information, which in turn assists the executive and managers make strategic decisions for the enterprise's benefit.

These assets could also include several record cubes, databases, or flat files. The statistical integration strategy is formally stated as a triple (**G, S, M**) approach.

G represents the **global** schema,

S represents the **heterogeneous** source of schema, and

M represents the **mapping** between source and global schema queries.

In this article, you will learn about Data integration in data mining and discuss its methods, issues, techniques, and tools.

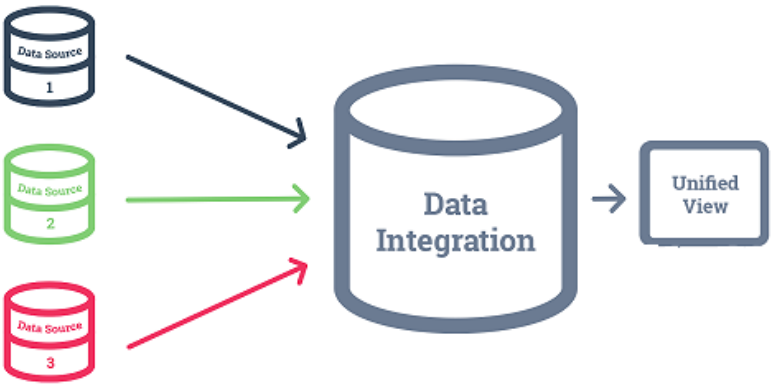


Figure 1: Data Integration Process

3.2 Why is the Data Integration Important?

Companies that want to stay competitive and relevant welcome big data and all of its benefits and drawbacks. One of the most common applications for data integration services and technologies is market and consumer data collection. Data integration supports queries in these vast datasets, benefiting from corporate intelligence and consumer data analytics to stimulate real-time information delivery. Enterprise data integration feeds integrated data into data centers to enable enterprise reporting, predictive analytics, and business intelligence.

Data integration is particularly important in the healthcare industry. Integrated data from various patient records and clinics assist clinicians in identifying medical disorders and diseases by integrating data from many systems into a single perspective of beneficial information from which useful insights can be derived. Effective data collection and integration also improve medical insurance claims processing accuracy and ensure that patient names and contact information are recorded consistently and accurately. Interoperability refers to the sharing of information across different systems.

3.3 Data Integration Approaches

There are mainly two types of approaches for data integration. These are as follows:

3.3.1 Tight Coupling

It is the process of using **ETL** (**Extraction, Transformation, and Loading**) to combine data from various sources into a single physical location.

3.3.2 Loose Coupling

Facts with loose coupling are most effectively kept in the actual source databases. This approach provides an interface that gets a query from the user, changes it into a format that the

supply database may understand, and then sends the query to the source databases without delay to obtain the result.

3.4 Issues in Data Integration

When you integrate the data in Data Mining, you may face many issues. There are some of those issues:

3.4.1 Entity Identification Problem

As you understand, the records are obtained from heterogeneous sources, and how can you 'match the real-world entities from the data'. For example, you were given client data from specialized statistics sites. Customer identity is assigned to an entity from one statistics supply, while a customer range is assigned to an entity from another statistics supply. Analyzing such metadata statistics will prevent you from making errors during schema integration.

Structural integration is completed by guaranteeing that the functional dependency and referential constraints of a character in the source machine match the functional dependency and referential constraints of the identical character in the target machine. For example, assume that the discount is applied to the entire order in one machine, but in every other machine, the discount is applied to each item in the order. This distinction should be noted before the information from those assets is included in the goal system.

3.4.2 Redundancy and Correlation Analysis

One of the major issues in the course of data integration is redundancy. Unimportant data that are no longer required are referred to as redundant data. It may also appear due to attributes created from the use of another property inside the information set. For example, if one truth set contains the patronage and distinct data set as the purchaser's date of the beginning, then age may be a redundant attribute because it can be deduced from the use of the beginning date.

Inconsistencies further increase the level of redundancy within the characteristic. The use of correlation analysis can be used to determine redundancy. The traits are examined to determine their interdependence on each difference, consequently discovering the link between them.

3.4.3 Tuple Duplication

Information integration has also handled duplicate tuples in addition to redundancy. Duplicate tuples may also appear in the generated information if the Denormalized table was utilized as a deliverable for data integration.

3.4.4 Data warfare Detection and backbone

The data warfare technique of combining records from several sources is unhealthy. In the same way, that characteristic values can vary, so can statistics units. The disparity may be related to the fact that they are represented differently within the special data units. For example, in one-of-a-kind towns, the price of an inn room might be expressed in a particular currency. This type of issue is recognized and fixed during the data integration process.

3.5 Data Integration Techniques

There are various data integration techniques in data mining. Some of them are as follows:

3.5.1 Middleware Integration

The middleware software is used to take data from many sources, normalize it, and store it in the resulting data set. When an enterprise needs to integrate data from legacy systems to modern systems, this technique is used. Middleware software acts as a translator between legacy and advanced systems. You may take an adapter that allows two systems with different interfaces to be connected. It is only applicable to certain systems.

3.5.2 Manual Integration

This method avoids using automation during data integration. The data analyst collects, cleans, and integrates the data to produce meaningful information. This strategy is suitable for a mini organization with a limited data set. Although, it will be time-consuming for the huge, sophisticated, and recurring integration. Because the entire process must be done manually, it is a time-consuming operation.

3.5.3 Application-based integration

It is using software applications to extract, transform, and load data from disparate sources. This strategy saves time and effort, but it is a little more complicated because building such an application necessitates technical understanding. This strategy saves time and effort, but it is a little more complicated because building such an application necessitates technical understanding.

3.5.4 Uniform Access Integration

This method combines data from a more disparate source. However, the data's position is not altered in this scenario; the data stays in its original location. This technique merely generates a unified view of the integrated data. The integrated data does not need to be stored separately because the end-user only sees the integrated view.

3.5.5 Data Warehousing

This technique is related to the uniform access integration technique in a roundabout way. The unified view, on the other hand, is stored in a different location. It enables the data analyst to deal with more sophisticated inquiries. Although it is a promising solution and increased storage costs, the unified data's view or copy requires separate storage and maintenance costs.

3.5.6 Integration tools

There are various integration tools in data mining. Some of them are as follows:

- **On-promise data integration tool**

An on-premise data integration tool integrates data from local sources and connects legacy databases using middleware software.

- **Open-source data integration tool**

If you want to avoid pricey enterprise solutions, an open-source data integration tool is the ideal alternative. Although, you will be responsible for the security and privacy of the data if you're using the tool.

- **Cloud-based data integration tool**

A cloud-based data integration tool may provide an 'integration platform as a service'.

CHAPTER – 4
DATA REDUCTION

4.1 Data Reduction in Data Mining

Data reduction techniques ensure the integrity of data while reducing the data. Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results.

Data reduction does not affect the result obtained from data mining. That means the result obtained from data mining before and after data reduction is the same or almost the same.

Data reduction aims to define it more compactly. When the data size is smaller, it is simpler to apply sophisticated and computationally high-priced algorithms. The reduction of the data may be in terms of the number of rows (records) or terms of the number of columns (dimensions).

4.2 Techniques of Data Reduction

Here are the following techniques or methods of data reduction in data mining, such as:

4.2.1 Dimensionality Reduction

Whenever we encounter weakly important data, we use the attribute required for our analysis. Dimensionality reduction eliminates the attributes from the data set under consideration, thereby reducing the volume of original data. It reduces data size as it eliminates outdated or redundant features. Here are three methods of dimensionality reduction.

4.2.2 Wavelet Transform

In the wavelet transform, suppose a data vector A is transformed into a numerically different data vector A' such that both A and A' vectors are of the same length. Then how it is useful in reducing data because the data obtained from the wavelet transform can be truncated. The compressed data is obtained by retaining the smallest fragment of the strongest wavelet coefficients. Wavelet transform can be applied to data cubes, sparse data, or skewed data.

4.2.3 Principal Component Analysis

Suppose we have a data set to be analyzed that has tuples with n attributes. The principal component analysis identifies k independent tuples with n attributes that can represent the data

set. In this way, the original data can be cast on a much smaller space, and dimensionality reduction can be achieved. Principal component analysis can be applied to sparse and skewed data.

4.2.4 Attribute Subset Selection

The large data set has many attributes, some of which are irrelevant to data mining or some are redundant. The core attribute subset selection reduces the data volume and dimensionality. The attribute subset selection reduces the volume of data by eliminating redundant and irrelevant attributes. The attribute subset selection ensures that we get a good subset of original attributes even after eliminating the unwanted attributes. The resulting probability of data distribution is as close as possible to the original data distribution using all the attributes.

4.2.5 Numeorsity Reduction

The numerosity reduction reduces the original data volume and represents it in a much smaller form. This technique includes two types parametric and non-parametric numerosity reduction.

- **Parametric**

Parametric numerosity reduction incorporates storing only data parameters instead of the original data. One method of parametric numerosity reduction is the regression and log-linear method.

Linear regression models a relationship between the two attributes by modelling a linear equation to the data set. Suppose we need to model a linear function between two attributes.

$y=wx+b$

Here, y is the response attribute, and x is the predictor attribute. If we discuss in terms of data mining, attribute x and attribute y are the numeric database attributes, whereas w and b are regression coefficients.

Multiple linear regressions let the response variable y model linear function between two or more predict or variables.

Log-linear model discovers the relation between two or more discrete attributes in the database. Suppose we have a set of tuples presented in n-dimensional space. Then the log-linear model is used to study the probability of each tuple in a multidimensional space. Regression and log-linear methods can be used for sparse data and skewed data.

- **Non-Parametric**

A non-parametric numerosity reduction technique does not assume any model. The non-Parametric technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the parametric. There are some Non-Parametric data reduction techniques, Histogram, Clustering, Sampling, etc.

- **Histogram**

A histogram is a graph that represents frequency distribution which describes how often a value appears in the data. Histogram uses the binning method to represent an attribute's data distribution. It uses a disjoint subset which we call bin or buckets.
A histogram can represent a dense, sparse, uniform, or skewed data. Instead of only one attribute, the histogram can be implemented for multiple attributes. It can effectively represent up to five attributes.

- **Clustering**

Clustering techniques groups similar objects from the data so that the objects in a cluster are similar to each other, but they are dissimilar to objects in another cluster.
How much similar are the objects inside a cluster can be calculated using a distance function. More is the similarity between the objects in a cluster closer they appear in the cluster.
The quality of the cluster depends on the diameter of the cluster, i.e., the max distance between any two objects in the cluster. The cluster representation replaces the original data. This technique is more effective if the present data can be classified into a distinct clustered.

- **Sampling**

One of the methods used for data reduction is sampling, as it can reduce the large data set into a much smaller data sample. Below we will discuss the different methods in which we can sample a large data set D containing N tuples:

Simple random sample without replacement (SRSWOR) of size s:

In this s, some tuples are drawn from N tuples such that in the data set D (s<N). The probability of drawing any tuple from the data set D is 1/N. This means all tuples have an equal probability of getting sampled.

Simple random sample with replacement (SRSWR) of size s:

It is similar to the SRSWOR, but the tuple is drawn from data set D, is recorded, and then replaced into the data set D so that it can be drawn again.

CHAPTER – 5

METHODS & METHODOLOGY

Data mining operations need a systematic approach. The process of Data mining is usually specified in the form of an ordered list but the process is not linear. Because if a particular time one may require to step back and rework on the previously perform steps.

The various steps of Data mining are:

5.1 Steps of Data mining Process

The various steps of Data mining process are as follows:

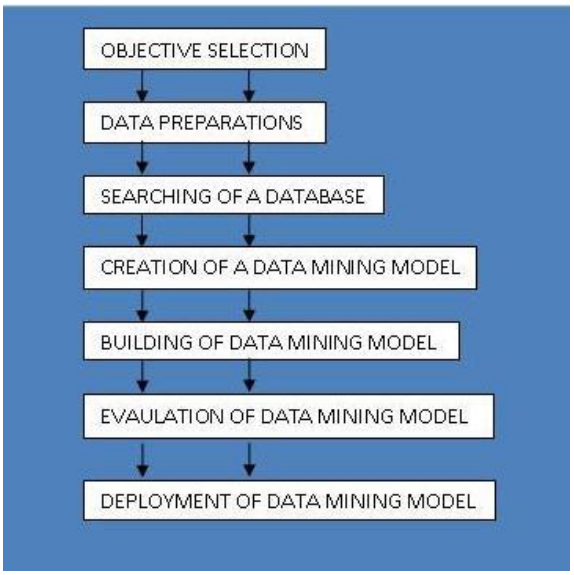


Figure 2: Main steps of Data mining Process

5.1.1 Objective Selection

The business objectives are decided prior to the Data mining process. These business objectives may depend on the inferences or results which will be drawn by Data mining processes. This can be achieved by joint effort of the data analyst who can translate the objectives identified by the analyst into a well-defined Data mining problem.

5.1.2 Data preparations

After deciding the objectives the data is prepared for the mining process. This process consists of the following sub steps:

- **Data Selection**

There are two parts to select data for Data mining. The first part, locating data tends to be more mechanical in nature than the second part. The second part identifying data requires

significant input by a domain expert for the data. A domain expert is one who is intimately familiar with the business purposes and aspects or domain of the data to be examined.

- **Data Cleaning**

Data cleaning is the process of ensuring that for Data mining purposes the data is uniform in terms of key and attributes usage. Data cleaning is separate data from data enrichment and data because data cleaning attempts to correct misused or in correct attributes in existing data. Data enrichment by contrast adds new attributes to existing data while data transformation changes to form or structure of attributes in existing data to meet specific Data mining requirements.

- **Data Enrichment**

Data enrichment is the process of adding new attributes such as calculated fields or data from external sources to existing data. This can include combining internal data with external data, obtained from either different departments or companies or vendors that sell standardized industry relevant data.

- **Data Transformation**

Data transformation is the process of changing the form or structure of existing attributes.

5.1.3 Searching of a Database

This phase is to select and examine important data sets of a Data mining database in order to determine their feasibility to solve the problem. Searching the database is a time consuming process and requires a good user interface and computer system with good processing speed.

5.1.4 Creation of Data mining Model

This phase is to select variables to act as predictors. New variables are also built depending upon the existing variables along with defining the range of variables in order to support imprecise information.

5.1.5 Building of Data mining Model

This phase is to create various Data mining models and to select the best of these models. Building a Data mining model is an iterative process. The Data mining model which we select can be a decision tree, an artificial neural network or an association rule model.

5.1.6 Evaluation of Data mining Model

This phase is to evaluate the accuracy of the selected Data mining model. In Data mining the evaluating parameter is data accuracy in order to test the working of the model. This is because the information generated in the simulated environment varies from the external environment.

5.1.7 Deployment of Data mining Model

This phase is to deploy the built and the evaluated Data mining model in the external working environment. A monitoring system should monitor the working of the model and produce reports about its performance. The information in the report helps to enhance the performance of selected Data mining model.

5.2 Data mining Process Models

We need to follow a systematic approach of Data mining for meaningful retrieval of data from large data banks. Several process models have been proposed by various individual and organizations that provides systematic steps for Data mining. The four most popular process models of Data mining are:

5.2.1 The 5A's process model

This process model stands for Assess, Access, Analyze, Act and Automate. The 5A's process model of Data mining generally begins by first assessing the problem in hand. The next logical step is to access or accumulate data that is related to the problem. After that we analyze the accumulated data from different analyses using various Data mining techniques. We then extract meaningful information from the analyzed data and implement the result in. solving the problem in hand. At least we try to automate the process of Data mining by building software that uses the various techniques which we used in the 5A's process model. The following Fig-2 shows the life cycle of the 5A's process model.

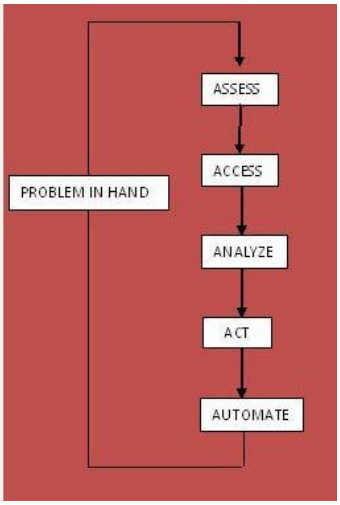


Figure 3: 5A' Process Model Life Cycle

5.2.2 The Crisp DM process model

In this process model Crisp DM stands, for cross industry standard process for Data mining. The life cycle of CRISP DM process model consists of six phases.

- **Understanding the business**

This phase is to understand the objectives and requirements of the business problem and generating a Data mining definition for the business problem.

- **Understanding the data**

This phase is to first analyze the data collected in the first phase and study its characteristics and matching patterns to propose a hypothesis for solving the problem.

- **Preparing the data**

This phase is to create final data sets that are input to various modeling tools. The raw data items are first transformed and cleaned to generate datasets which are in the form of tables, records and fields.

- **Modeling**

This phase is to select and apply different modeling techniques of Data mining. We input the data sets collected from the provides phase to these modeling techniques and analyze the generated output.

- **Evaluation**

This phase is to evaluate a model or a set of models that you generate in the previous phase for better analysis of the refined data.

- **Deployment**

This phase is to organize and impresent the knowledge gained from the evaluation phase in such a way that it is easy for the end users to compare hand.

The following Fig-4 shows the life cycle of the CRISP DM process model.

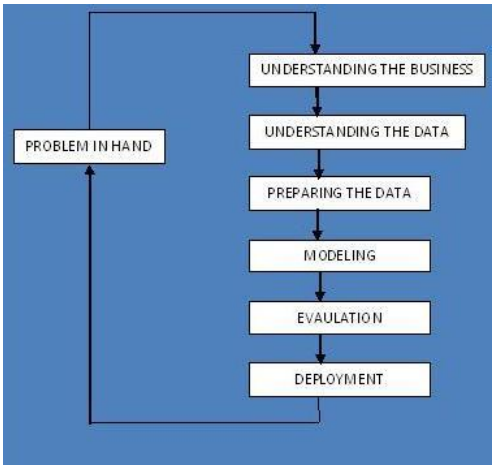


Figure 4: CRISP DM Process Model

5.2.3 The SEMAA Process Model

In this process model SEEMA stands for sample, Explore, Modify, Modal and Assess.The life cycle of the SEMAA process model consists of five phases.

- **Sample**

This phase is to extract a portion from a large data bank such that we are able to retrieve meaningful information from the extracted portion of data.

- **Explore**

This phase is to explore and refine the sample portion of data using various statistical Data mining techniques in order to search for unusual trends and irregularities in the sample data.

- **Modify**

This phase is to modify the explored data by creating; selecting and transforming the predictive variables for the selection of a prospective Data mining model. As per the problem in hand we may need to add new predictive variables or delete existing predictive variables to narrow down the search for a useful solution to the problem.

- **Model**

This phase is to select a Data mining model that automatically searches for a combination of data which we can use to predict the requirement result for the problem. Some of the modeling technique that we can use as a model is neural network and statistical modules.

- **Asses**

This phase is to assess the use and reliability of the data generated by the model that we selected in the previous phase and estimate its performance.

The following Fig-5 shows the life cycle of the SEEMMA process model.

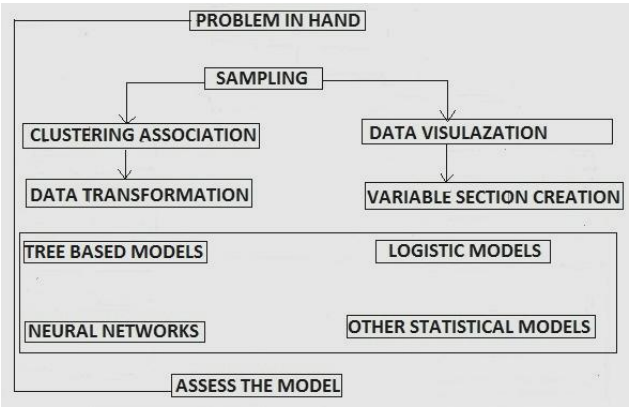


Figure 5: SEMMA Process Model

5.2.4 The Six Sigma Process Model

The six sigma process is a data driven process model that eliminates defects, wastes or quality control problems that generally occurs in a production environment. Six Sigma is very popular in various American Industries due to its early implementation and it is likely to be implemented worldwide. This process model is based on various statistical techniques, use of various types of data analysis techniques and implementation of systematic training of all the employees of an organization. Six sigma process model postulates a sequence of five stages called DMAIC, which stands for Define, measure, Analyze, Improve and control.

The life cycle of six sigma process model consists of five phases:

- **Define:** This phase is to define the goals of project along with its limitations.
- **Measure:** This phase is to collect information about the current process in which the work in done and to try to identify the basics of the problem.
- **Analyze:** This phase is to identify the root cause of the problem in hand and insure those root causes by using various data analysis tools.
- **Improve:** This phase is to implement all these solutions that tricks and solves the root causes of the problem in hand.
- **Control:** This phase is to monitor the outcome of all its previous phases and suggest improvement measures in each of its earlier phases.

The following Fig-6 shows the life cycle of the Six Sigma process model.

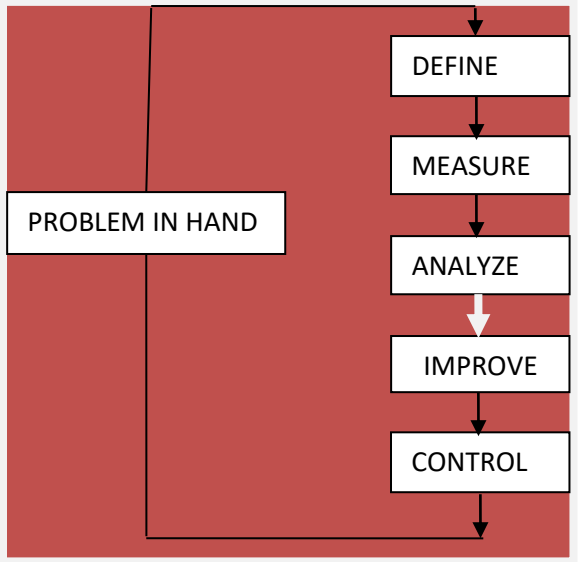


Figure 6: Six Sigma Process Model

CHAPTER – 6
DATA MINING TECHNIQUES

6.1 Introduction of Data Mining Techniques

In Data mining a decision tree is a predictive model which can be used to both classifiers and regression models. In operational research on the other hand decision tree refer to a hierarchical model of decisions and their consequences most likely t reach the goal.

When a decision tree s used for classification tasks, it is more appropriately referred to as a classifier tree and when it is used for regression tasks. It is called a regression tree. Here we concentrate mainly on classification trees. Classification trees are used to classify an object or an instance (such as insurant) to a predefined set of classes (such as risky/non risky) based on their attributes values (such as age or gender), classification trees are frequently used in applied fields such as finance, marketing, engineering and medicine. The classification tree is useful as an explore at ordinary technician. However it does not attempt there are many other techniques which can be used to classify or predict the membership of instances to a predefined set of classes such as artificial neural networks or support vector machines.

The use of a decision tree is a very popular technique in Data mining. In the opinion of many reporters decision trees are popular due to their simplicity and transparency.

Decision trees are self explanatory; there is no need to be a Data mining expert in order to follow a certain decision tree. Classification trees are usually represented graphically as hierarchical structures making them easier to interpret than other techniques. If the classification tree process complicated (i.e. has many nodes) then its straight forward graphical representation becomes useless. For complex trees other graphical procedures should be developed to simplify interpretation.

Tree size: Decision makes prefer a decision tree that is not complex since it is apt to be more comprehensive. Furthermore according to Breiman *et al.* (1984) trees compulsorily have a crucial effect or it’s accurately.

Usually the tree complexity is measured by one of the following matrices: the total number of models, total number of learns, tree depth and number of attributes used. The tree complexity is explicitly controlled by the stopping criteria and the pruning method that are employed.

6.2 Types of Data Mining Techniques:

Various Data mining techniques are as follows:

- Association Rule
- Clustering Techniques
- Genetic Algorithm
- Artificial Neural Network
- Decision Tree

6.2.1 Association Rule

Association rule Data mining techniques searches an interesting relationship among items in a given data set. It finds interesting correlation relationships or association among large sets of given items. Association rules show attributes value conditions which occur frequently together in a given dataset.

A typical and widely used example of association rule mining is market basket analysis.

Market basket

Analysis is a modeling technique based on the theory that if you buy ascertain group of items then you are more (or less) likely to buy another group of items. The set of items a customer buys is referred to as an item set and market basket analysis finds to discover relationship between purchases.

The relationship will be in the form of a rule:

IF {Bread} THEN {Butter}

The above condition extraction extracts the hidden information 1.c if a customer used to buy Bread he will also buy Butter as side dish.

There are two types of association rule levels.

- Support level
- Confidence level

Rules of support level

Let T be the set of all transaction under consideration e.g. let T be the set of all “Baskets” or Carts a products bought by the customers from a supermarket, say on given day. The support of an item set S is the percentage of those transactions in T which contain Sin the supermarket example this is the number of baskets that contain a given set S of products for example S={Bread, Butter, Milk},If U is the set of as transactions which contain all items in S then

Support (S) = ($\left|U\right|$ / $\left|T\right|$)*100%

Where |U| & |T| are the number of elements in U and T respectively. For example if a customer buys the set X={Milk, Bread, Apples, banana, Samsosa, Cheese, Onions, Potatoes} then S is obviously a subset of X and hence S is in U.If there are 318 customers and 242 of them buy such a set U or a similar one that contains S then support (S)=(242/318)=76.1%

6.2.2 Clustering Techniques

The process of organizing objects into groups whose members are similar in some way is called clustering. Thus cluster is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. There are two types of clustering.

- **Distance based Clustering**

In this type of clustering the similarity criterion is distance i.e. or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance).

- **Conceptual Clustering**

In this type of clustering two or more objects belong to the same cluster if this one defines a concept common to all that objects.In other words objects are grouped according to their fit to description concepts not according to simple similarity measures.

Given two p dimensional data objects $i=(x_{i1},x_{i2},x_{i3},\ldots.x_{ip})$ and $j=(x_{j1},x_{j2},xj_3\ldots..x_{jp})$.The following common distance functions can be defined:

Euclidean distance function

$$d(i,j)=\sqrt{\left|(x_{i1}-x_{j1})\right|^2+\left|(x_{i2}-x_{j2})\right|^2+\cdots+\left|(x_{ip}-x_{jp})\right|^2}$$

Manhattan distance function

$$d(i,j)=\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|\ldots..+\left|x_{ip}-x_{jp}\right|$$

When using the Euclidean distance functions to compare distance it is not necessary to calculate the square root because distance are always positive numbers and assume for two distance d_1 and d2, $>\sqrt{d1}>\sqrt{d2}$ d1>d2.If same of an objects attributes are measured along different scale so when using the Euclidean distance functions, attributes with larger scales of measurement may overwhelm attributes measured on a smaller scale. To prevent this problem the attribute values are often normalized to lie between 0 and 1.

- **Clustering Algorithms**

Clustering techniques: according to Han J kamber cluster techniques is the process of portioning data objects (records, documents etc) into meaningful groups or clusters so that

objects within a cluster have similar characteristics but are dissimilar to objects in other cluster. Clustering can be understood as unsupervised classification of unlabeled patterns (observations, data items or feature vectors) because no predefined category labels are associated with the objects in the training dataset. Clustering of web documents viewed by internet users can shows collections of documents belonging to the same topic.

Clustering has also been with mankind since very beginning.people cluster tighter according to their certain characteristics, qualities, and attributes people from the same country religion tribe race etc. Cluster together. (Wires et al 2010) have pointed out that data clustering allows us to construct simpler understand able modules of that world which can be worked upon more comfortably.

Clustering in another Data mining technique which can because to detect the terrorism and crime. Classification and clustering are almost the same but whereas classification requires basis parameter but clustering does not require any parameter. Clustering techniques and algorithm are dependent on real life model which individual with some virtues must cluster together.

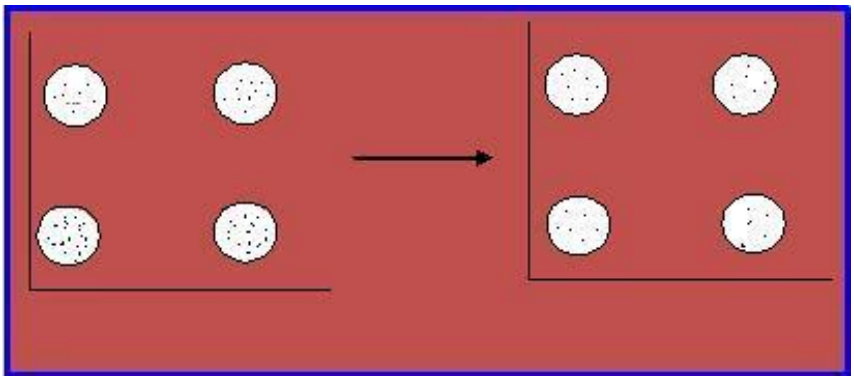


Figure 7: Clustering Process

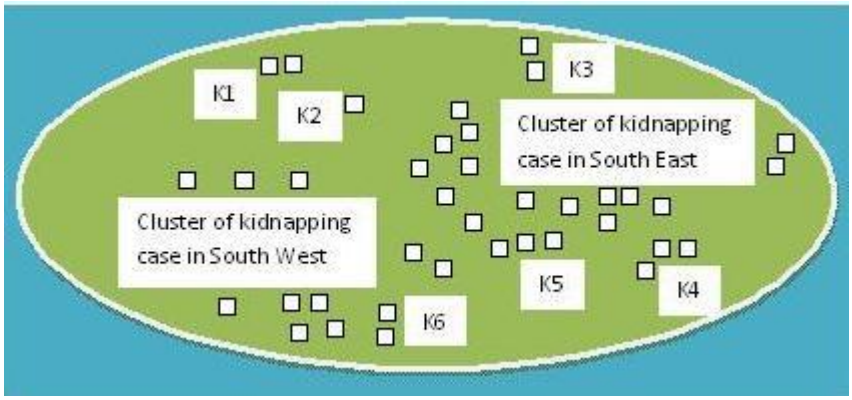


Figure 8: Cluster Cases

Fig-8 shows a sample of cluster cases of kidnapping in some selected areas. Cluster assumes that in crime dominates area; individuals with some crime specialties will cluster together.

For example individuals who are specialties in kidnapping tend to cluster together. Clustering technique will then identify given cluster and their areas of operation any time when a crime is reports. Then the law enforcement agencies can observe the related clusters and examine them for clues.

Two crows has defined that clustering is a way to segregate data into groups which are not previously defined whereas classification is a way to segment data by assigning it to groups that are already defined.

Fig-8 represents a typical clustering algorithm, here the data set is clustered into various clusters, and then crime suspects are placed under each cluster for comparisons. If they tally then they are placed under surveillance a new cluster is then formed and the dataset updated.

Sometimes a clustering anomaly occurs and is referenced anomaly detection. This is the case of a relevant event which happens within a particular place and an event activity happening elsewhere: Fig- shows that when an anomaly occurs, it also shows the starting of another cluster. In the following fig the normal clusters of kidnapping cases whereas l2, l4, l5 and l6 are all anomalies of kidnapping.

K-Nearest Neighbor (K-NN):

According to two crows (1999) K-Nearest Neighbor (K-NN) is a classification technique that decides in which class to place a new case by examining some number the “k” in k nearest neighbor of the most similar cases or neighbor. It counts the number of cases for each class and then assigns the new case to the same class to which most of its neighbors belong. This technique and its algorithm have been used for ages. When one observe a decent person who becomes close to suspected criminals then by intuition hear she will carriage this technique power forms very well when we have identified a group, the chances occur that any other person among the group will likely be associated with them. Let us take an example of a group of notorious armed bandits. When any is seen near each member of the group our perception is that hear she is one, another example is that in a place that is notorious forits gang activities then if any person is seen wondering around he or she will automatically be assumed to be a member of that gang.

6.2.3 Genetic Algorithm

Optimization techniques that use process such as genetic combination, mutation and natural selection in a design based on the concepts of evolution.

Genetic algorithm is based on Darwin theory of survival of the fittest. Here the best program or logic survives from a pool of solutions. Two programs called chromosomes combine to produce a third program called Chaid, the reproduction process goes through crossover and mutation operations.

Crossover

Applying simple point crossover reproduction mechanisms appoint along the chromosome length is randomly chosen at which the crossover is done as sown in following fig.

- **Single point crossover**

In this type one crossover point is chosen and the string from the beginning of chromosome to the crossover point in copied from one parent and the rest is copied from the second parent resulting in a child.We consider the following chromosome and crossover point at position 4.

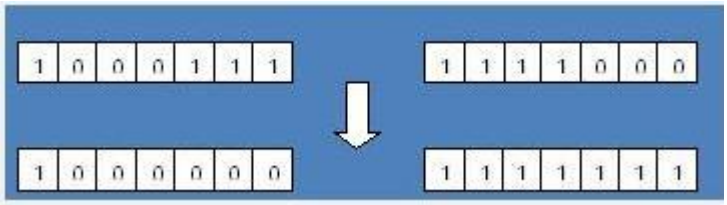


Figure 9: Single point crossover

We see here that the regions after the crossover point are interchanged in the children.

- **Two point crossovers**

In this type two crossover points are selected and the string from beginning of one chromosome to the first crossover point is copied from one parent, the past from the first to the second crossover point is copied from the second parent and the rest is copied from the first parent. This type of crossover is mainly employed in permutation encoding and value encoding where a single, point crossover would result in consistencies in the client chromosomes. We consider the following chromosomes and crossover points at positions 2 and 5. We see that the crossover requires between the crossovers points are interchanged in the children.

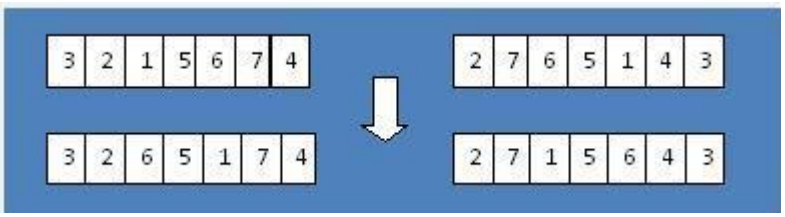


Figure 10: Two point crossover

- **Tree crossover**

The method of tree crossover is most suitable when tree encoding is employed. One crossover point is chosen at random and parents are divided in that point and parts below crossover point are exchanged to produce new offering.

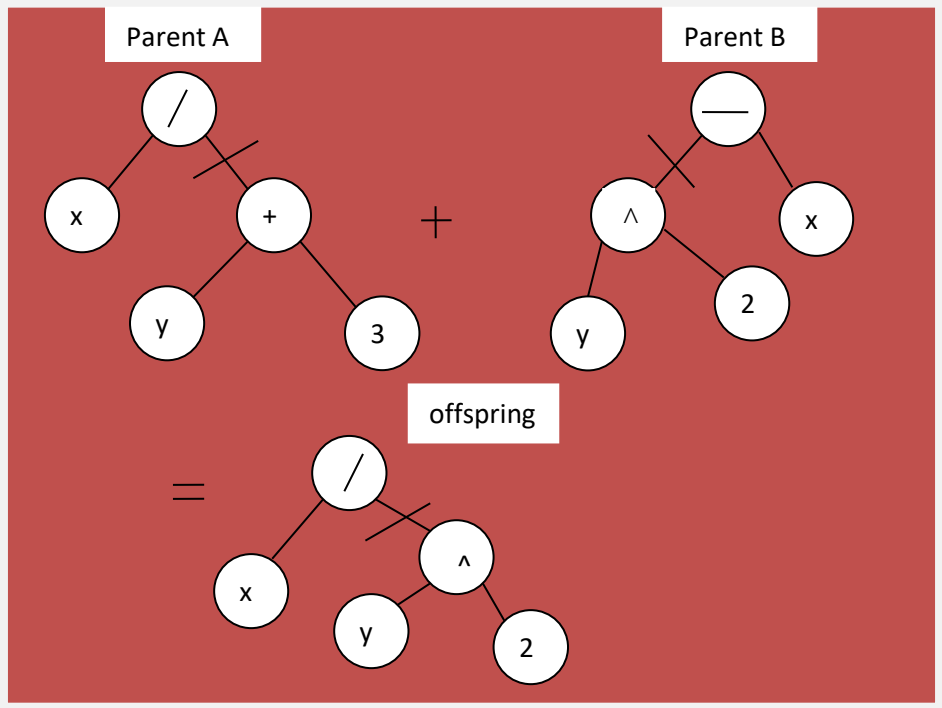


Figure 11: Tree Crossover

Rules of confidence level

If "A then B" rule confidence is the conditional probability that B is true when A is known to be true.

To evaluate association rules the confidence of rule R=A and B→C" is the support of the set of all items which appear in the rule divided by the support of the antecedent of the rule. Confidence (R) = (support ({A, B, C})/support ({A, B}))*100.

More intensively the confidence of a rule is the number of cases in which the rule is correct relative to the number of cases in which it is applicable. For example let R= "Butter and Bread→ Milk".

If a customer buys Butter and Bread then the rule is applicable and it says that he/she can be expected to buy Milk. If he/she doses not buy sugar or does not buy Bread or buys neither than the rule is not applicable nor these does not say anything about this customer.

6.2.4 Artificial Neural Network

Neural network are analytic techniques modeled after the (hypothesized) process of learning in the cognitive system and the neurons logical functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after other observations (on the same or other variables) after executing a process of so called learning from existing data.

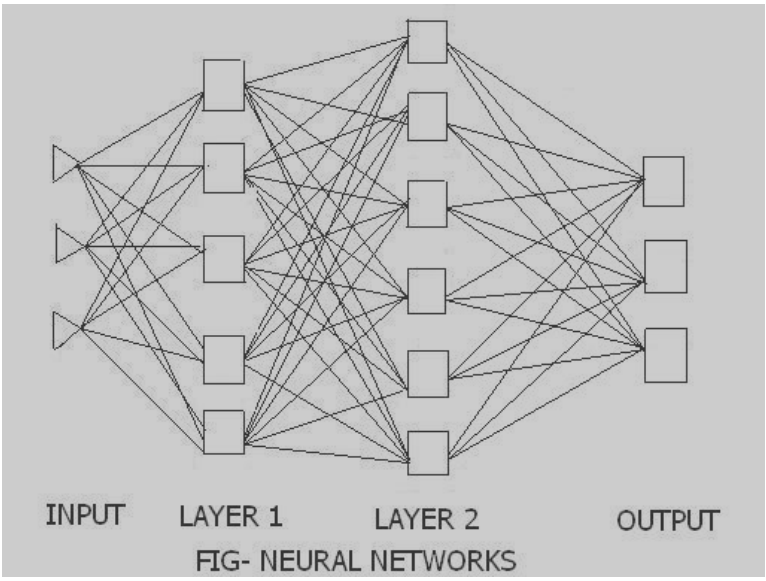


Figure 12: Neural Network

The first step is to design specific network architecture (that includes a specific number of layers, each consisting of a certain number of "neurons".

The size and structure of the networks needs to match the nature (e.g. the formal complexity) of the investigated phenomenon. Because the latter is obviously not known very well at this early stage, this task is not easy and often involves multiple "trails and errors". (Now there is however neural network software that applies artificial intelligence techniques to aid in that tedious task and finds "the best" network architecture. The new network is then subjected to the process of "training". In that phase, neurons apply an iterative process to the number of

inputs (variables) to adjust the weights of the network in order to optimally predict (in traditional terms one could say, find a "fit" to) the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions.

Neural Networks techniques can be used as a component of Analyses designed to build explanatory models because It can help explore data sets in search for relevant variables or groups of Variables.

6.2.5 Decision Tree

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

6.2.5.1 Introduction

A decision tree is a flow chart like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution. The topmost node in a tree is the root node. A decision tree can also be defined as a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

Decision tree represents rules, so decision tree is a classifier in the form of a tree structure where each node is either,

A leaf node, indicating a class of instances or

A decision node that specifies some text to be carried out on a single attribute value, with one branch and sub tree for each possible outcome of the test.

In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree.

A path is traced from the root to a leaf node which holds the class prediction for that sample. Decision trees can easily be converted to classification rules.

The construction of decision tree classifier does not require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery.

The representation of acquired knowledge in tree form is spontaneous and generally easy to understand by humans. The learning and classification steps of decision tree induction are simpler and fast. In general decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing and production, financial analysis, astronomy and molecular biology. Decision trees are the basis of several commercial rule induction systems.

The key requirements to do mining with decision trees are:

- **Attribute value description:** Object must be expressible in terms of a fixed collection of properties or attributes.
- **Predefined classes:** The categories to which example are to be assigned must have been established beforehand(supervised data)
- **Discrete classes:** A class does or does not belong to a particular class and there must be more cases than classes.
- **Sufficient data:** usually hundreds or even thousands of training cases.

6.2.5.2 Attributes Selection Measure:

Splitting Indices

In this section we shall study two different methods of determining the goodness of split. One is index based on the information theory that is information gain based on entropy. The other one is derived from economics as measure of diversity. This is called the gini index.

The Information Gain Measure

It is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split**.** The attribute with highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or “impurity”. In there partitions such an information theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

Let S be a set consisting of S data samples. Suppose the class label attribute has m distinct values defining m distance classes, C_i (for i=1 …., m).let Si is the number of samples of in class C_i the expected information needed to classify a given sample is given by:

$$\sum_{i=1}^{m} p_i \log_2 p_i I(S_1,S_2\ldots..S_m) = -$$

Where pi is the probability that an arbitrary sample belongs to class C_i and is estimated by sells. Note that a log function to the base 2 is used since the information is encoded in bits.

Let attribute A have V distinct values, $\{a_1, a_2,.a_v\}$.Attribute A can be used to partition S into V subsets $\{S_1, S_2\ldots.S_v\}$, where S_j contains those samples in S that have value a_j of A. If A were selected as the test attribute (i.e.,best attribute for splitting),then there subsets would correspond to the branches grown from the node containing the set S. let S_{ij} be the number of samples of class C_i in a subset S_j.The entropy or expected information based on the partitioning into subsets by A is given by:

$$\sum_{j=1}^{m} Sij + \cdots..Smj/S$$

E (A) = I ($S_{1j}\ldots S_{mj}$)

The term $S_{1j}+\ldots.S_{mj}/S$ acts as the weight of j^{th} subset. and is the number of samples in the subset (i.e., having value a_j of A)divided by the total number of samples in S. The smaller the entropy value is, the greater the purity of the subset partitions. Note that for a given subset, Sj.

$$\sum_{i=1}^{m} P_{ij \log_2(P_{ij})} I\ (S_{1j}, S_{2j}\ldots.S_{mj}) = -$$

Where $P_{ij} = S_{ij}/S_j$

And is the probability a sample in Si belongs to class C_i.The encoding information that would be gained by branching on A is

Gain (A) =I (S_1, $S_{2\ldots}$, S_m)-E (A)

In other words, gain (A) is the expected reduction in entropy caused by knowing the value of attribute A.

The algorithm computes the information gain of each attribute the attribute with the highest information gain is chosen as the test attribute for the given set S.A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

In abstract, decision tree induction algorithms have been used for classification in a wide range of application domains. Such systems do not use domain knowledge, the learning and classification steps of decision tree induction are normally fast. Classification accuracy is typically high for data where the mapping of classes consists of long and then regions in concept space.

6.2.5.3 Construction of Decision Tree

An Example of construction of decision tree is given ahead:

The weather attributes are Outlook, Temperature, humidity and wind speed. They can have the following values:

Outlook = {Sunny, Overcast, Rain}

Temperature = {Hot, Mild, Cold}

Humidity = {High, Normal}

Wind = {Weak, Strong}

Table 1:

Day	Outlook	Temperature	Humidity	Wind	Play Table Tenis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cold	Normal	Weak	Yes
D6	Rain	Cold	Normal	Strong	No
D7	Overcast	Cold	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Construct decision tree classification algorithms: Decide on which day you can Play TABLE TENIS.

Attribute<Play TABLE TENIS> has 2 values {Yes, No}

If S is a collection of 14 examples with 9 Yes and 5 No examples then,

Info Gained before splitting I e Entropy(S)

$I(S_1, S_2) = -\log_2 (Si/S)$

$I(9, 5) = -(9/14)\log_2(9/14)-(5/14)\log_2(5/14)$

$= 0.4097+0.5305=0.9402$

Information gain is based on to decrease in entropy after a dataset is split on an attribute looking for which attribute creates the most homogeneous branches.

Outlook:

S_{11} = Sunny + Play TABLE TENIS (Yes) (2)

S_{21} = Sunny + Play TABLE TENIS (No) (3)

S_{12} = Overcast + Play TABLE TENIS (Yes) (4)

S_{22} = Overcast + Play TABLE TENIS (No) (0)

S_{13} = Rain + Play TABLE TENIS (Yes) (3)

S_{23} = Rain + Play TABLE TENIS (No) (2)

E(Outlook) = (S/14)I(S11,S21)+(4/14)IS12,S22)+(5/14)I(S13,S23)

$I(S_{11}, S_{21}) = -\log_2 (Si/s)$

$= -(2/5)\log_2(2/50-(3/5)\log_2(3/5)$

$= 0.9708$

$I(S_{12}, S_{22}) = -\log_2 (Si/5)$

$= -(4/4)\log_2(4/4) +0=0$

$I(S_{13}, S_{23}) = -\log_2 (Si/5)$

$= -(2/5)\log_2(2/5)-(3/5)\log_2(3/5)$

$= 0.9708$

So,

E(Outlook) = (S/14)I(S11,S25)+(4/14)I(S12,S22)+(S/14)I(S13,S23)

$= (5/14)*(0.9708) + (4/14)*(0) + (5/14)*(0.9708)$

$= 0.6930$

Gain (Outlook) $= I(S_1, S_2)$-E (Outlook)

$= 0.9406-0.693$

= 0.2472

Temperature:

S_{11} = Hot + Play TABLE TENIS (Yes) (2)

S_{21} = Hot + Play TABLE TENIS (No) (2)

S_{12} = Mild + Play TABLE TENIS (Yes) (4)

S_{22} = Mild + Play TABLE TENIS (No) (2)

S_{13} = Cold + Play TABLE TENIS (Yes) (3)

S_{23} = Cold + Play TABLE TENIS (No) (1)

$I(S_{11}, S_{21}) = - \log_2 (Si/S)$

$= - (2/4) \log_2 (2/4) - (2/4) \log_2 (2/4)$

= 1

$I(S_{12}, S_{22}) = - \log_2 (Si/S)$

$= - (3/4) \log_2 (3/4) - (1/4) \log_2 (1/4)$

= 0.811

So, E (Temperature)

= (S/14)I(S11,S21)+(4/14)I(S12,S22)+(S/14)I(S13,S23)

= (4/14)*(1) + (6/14)*(0.9179) + (4/14)*(0.811)

= 0.9108

Gain (Temperature) $= I(S_1, S_2)$-E (Temperature)

= 0.9406-0.9108

= 0.029

Humidity:

S_{11} = High + Play TABLE TENIS (Yes) (3)

S_{21} = High + Play TABLE TENIS (no) (4)

S_{12} = Normal + Play TABLE TENIS (Yes) (6)

S_{22} = Normal + Play TABLE TENIS (No) (1)

E (Humidity) = (7/14) $I(S_{11}, S_{21})$ + (7/14) $I(S_{12}, S_{22})$

$I(S_{11}, S_{21}) = - \log_2 (Si/S)$

$= - (3/7) \log_2 (3/7) - (4/7) \log_2 (4/7)$

= 0.523 +0.4613

= 0.9843

$I(S_{12}, S_{22}) = - \log_2 (Si/S)$

$= - (6/7) \log_2 (6/7) - (1/7) \log_2 (6/7)$

= 0.5916

So,

E (Humidity) = (7/14)I(S11,S21)+(7/14)I(S_{12},S_{22})

= (7/14)*(0.9843) + (7/14)*(0.5916)

= 0.7879

Gain (Humidity) = 0.9402-0.7879

= 0.152

Wind:

S_{11} = Weak + Play TABLE TENIS (Yes) (6)

S_{21} = Weak + Play TABLE TENIS (No) (2)

S_{12} = Strong + Play TABLE TENIS (Yes) (3)

S_{22} = Strong + Play TABLE TENIS (No) (3)

I (S_{11}, S_{21}) = - $\log_2$ (Si/S)

= - (6/8) $\log_2$ (6/8)-(2/8) $\log_2$ (2/8)

= 0.311+ (1/2)

= 0.811

I (S_{12}, S_{22}) = - $\log_2$ (Si/S)

= - (3/6) $\log_2$ (3/6)-(3/6) $\log_2$ (3/6) =1

So, E (Wind) = (8/14) I (S_{11},S_{21})+(6/14)I(S_{12},S_{22})=0.892

Gain (Wind) = 0.9402-0.892=0.0486

Outlook attribute has the highest gain, therefore it is used as the decision attribute in the root node.

Since Outlook has three possible values, the root node has three branches (Sunny, Overcast, and Rain).

The next questions is “what attribute should be tested at the sunny branch node?”Since we have used Outlook at the root, we only decide on the remaining three attributes: Humidity, temperature or Wind.

Sunny= {D1, D2, D8, D9, D11} =5 with Outlook=Sunny

Gain (Sunny, Humidity) =0.970

Gain (Sunny, Temperature) =0.570

Gain (Sunny, Wind) =0.019

Humidity has the highest gain; therefore, it is used as the decision node. This process goes on until all data is classified perfectly or we run out of attributes.

The final decision tree will be

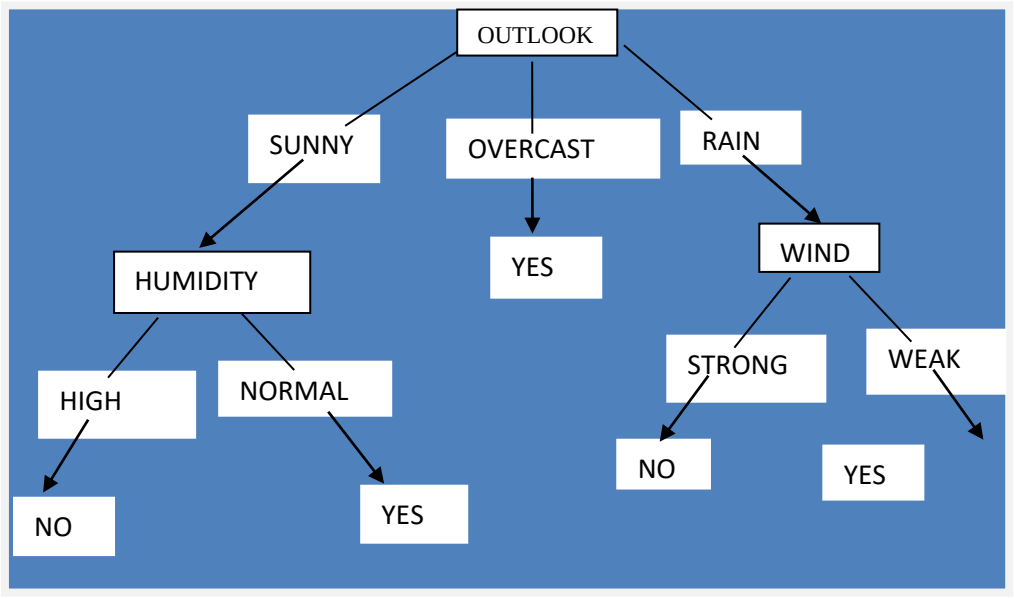


Figure 13: Decision tree for Game

6.2.5.4 Classification Rules from Decision Tree

The decision tree can also be expressed in rule format:

If Outlook=Sunny AND Humidity=High THEN Play TABLE TENIS=No

If outlook=Sunny AND Humidity=Normal THEN Play TABLE TENIS=Yes

If outlook=Rain AND Wind = Strong THEN play TABLE TENIS=No

If Outlook=Overcast THEN Play TABLE TENIS=Yes

If Outlook=Rain AND wind =Weak THEN Play TABLE TENIS=Yes

6.2.7 Decision Tree & Algorithms

The name algorithm' is given after the name of Abu Jafar Muhammad ibn Musa Alkl Warizmi,9th century is defined as follows:

An algorithm is a set of rules for carrying out a calculation either by hand or on a machine.

An algorithm is a sequence of computational steps that transform the input into the output.

An algorithm is a sequence of operations performed on data that have to be organized in data structures.

A finite set of instruction that specify a sequence of operations to be carried out in order to solve a specific problem or class of problems is called an algorithm.

An algorithm is an abstraction of a program to be executed on a physical machine (Model of Computation)

An algorithm is a set of rules that must be followed when solving a specific problem. Algorithm can also be defined as a well-defined computational procedure which takes some value or set of values as input and generates some value or set of values as output.

6.2.8 Types of Algorithm It is classified in many ways:

6.2.8.1 Cart

It was developed by Breimarn et al (1994). Cart Classification and regression Tree is one of the popular methods of building decision tree in the machine learning community. CART builds a binary decision tree by splitting the records at each node; according to a function of a single attribute. CART uses the Gini index for determining the best split. CART follows the above principle of constructing the decision tree. We outline the method for the sake of completeness.

The initial split produces two nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine all the input fields to find the candidate splitters. If no split can be found that significantly decreases the diversity of a given node, we label of it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of over fitting.

At the end of the tree growing process, every record as the training set has been assigned process, every record as the training set has been assigned to some leaf as the full decision tree, and each leaf can now be assigned a class and an error rate. The error rate as a leaf node is the percentage as incorrect classification at that node. The error rate as an entire decision tree is a weighted saws as the error rates as all rate at that leaf's contribution to the total is the error rate at that leaf multiplied by the probability that a record will end up in there.

6.2.8.2 ID3

It was developed by Quinlan (1986). Quinlan introduced the ID3, Iterative Dichotomizer3, for constructing the decision trees from data. In ID3, each node corresponds to splitting attribute and each are is a possible value of that attribute. At each node the splitting attribute is selected to be the most informative among the attributes not yet considered in the path

from the root. Entropy is used to measure how informative is node this algorithm use the criterion of information gain to determine the goodness of s the attributes with the greatest information gain is taken for all distinct values of the attributes.

6.2.8.3 C4.5

It was also developed by Quinlan (1993).C4.5 is an extension of ID3 that accounts unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. In building a decision trees we can deal with training sets that have records with unknown attributes values by evaluating the gain, or the gain ratio, for an attribute by considering only those records where those attributes values are available.We can classify records that have unknown attributes values by estimating the probability of the various possible results. Unlike CART, which generates a binary decision tree, C4.5 produces tree with variable branches per node. When a discreet variable is chosen as the splitting attributes in C4.5, there will be one branch for each value of the attributes.

6.2.8.4 CHAID

CHAID, proposed by KASS in 1980,is a derivative of AID(Automatic Interaction Detection), proposed by Hartigan in 1975.CHAID attempts to stop growing the tree before over fitting occurs, whereas the above algorithms generate a fully grown tree and the carry out pruning as past-processing step. In that sense, CHAID avoids the burning phase.

In the standard manner, the decision tree is constructed by partitioning the data set into two or more subsets, based on the values of one of the non class attributes. After the data set is partitioned according to the chosen attributes, each subset is considered for further partitioning using the same algorithm. Each subset is portioned without regard to any other subset; this process is repeated for each subset until some stopping criterion is met. In CHAID, the numbers of subsets in a partition can change from two up to the number of distinct values of the splitting attribute. In this regard, CHAID differs from CART, which always forms binary splits, and from ID3 or C4.5, which form a branch for every distinct value. The splitting attribute is chosen as the one that is most significantly associated with the dependent attributes according to a Chi-squared test as independence in a contingency table (across tabulation on he non class and class attribute). The main stopping criterion used by such methods is the p value from this Chi-squared test. A small p value indicates that the observed association between the splitting attributes and the dependent variable is unlikely to have occurred solely as the result on sampling variability.

If a splitting attribute has more that two possible values then there may be a way large number of ways to partition the data set based on these values. A combinational search algorithm

can be used to find a portions' that has a small p- value for the chi- square test. The p- values for each chi- squared test are adjusted for the multiplicity on partition. A bonafide adjustment is used for the p value computed from the contingency tables, relating the predictors to the dependent variable. The adjustment is conditional on the number of branches (compound categories) in the partition and thus not takes into account the fact that different numbers on branches are considered.

6.2.9 Strength & Weakness of Decision Tree Method

The strengths of decision tree methods are:

- Decision trees perform classification without requiring much computation.
- Decision trees are able to generate understandable rules.
- Decision trees are able to deal both continuous and categorical variables
- Decision trees provide a clear picture of which fields are most important for prediction or classifications.

The weaknesses of decision tree methods are:

- Decision trees do not treat well non rectangular reasons.
- Decision trees are less appropriate for estimation tasks, where the goal is to predict the value of a continuous attributes.
- The process of growing a decision tree is computationally expensive.
- Decision trees are prone to errors in classification problems with many class and relatively small no of training examples.

CHAPTER – 7
APPLICATIONS OF DATA MINING

7.1 Data Mining In Marketing

There are various applications which are more useful for data mining. Such as Data Mining In Marketing, Data Mining In Healthcare, Data Mining In Terror Related Activity, Defense Sector Transportation etc. Some of them are described below:

7.1.1 Introduction

Now a day Industries collect huge volumes of data on a daily basis. To analyze this data and discover meaningful information contained by it became an essential need for business. As the business environment develops and changes constantly facing every day new challenges the industries try to strengthen their market position and achieve competitive advantage by using new and innovative solutions like Data mining. Data mining is simply filtering through large amounts of raw data for useful information that gives business a competition edge.

At present, the business operation model has gradually turned from product focused to customer centric. Enterprises have to come to realize that customers' information is one of their key assets. As the enterprises explore the customers behavior in depth, they find that not all customers will bouncing profits adjust a small percentage of all users of the products the best customers account for a large portion of an organizations sales. Since customers are different hence concentrating on the heavy user market segment is an attractive strategy.

The customer data of an organization is collected from the interaction with customer such as customer's basic data and the sales transaction data. By analyzing the data the organization can understand customer differences. Once organization learns more and more about their customers they can use that knowledge to serve then better.

A Customer Relationship Management (CRM) system is a process to compile information that increases understanding of how to manage an organizations relationship with its costumers (Zikmund *et al.* , 2003).By cooperating with the marketing activities CRM system can bring a lot of profit and health enterprises to survive in a continuous changing and competitive environment.Most business agree that it cast many time(about six times),as much to get a new customers as it does to keep and old customer (Prahlad *et al.* , 2000) have shown that if an

organization can increased 5% of customer retention rate, profits from customers will move up 25% on average.

Data mining is a well-known technique that can be used to turn customer data into customer knowledge. Typically, 20% of the customers buy 80% of the product sold and it is the famous 80/20 principle. These 20% are heavy users and may be the best customers. In terms of marketing, focuses on heavy users can get more revenues.

Therefore, organizations should divide a heterogeneous market into a number of smaller more homogeneous subgroups and that is called computer segmentation. If an organization could make the marketing decision for new coming costumers according to their basic data it can avoid unnecessary marketing cost and complex Data mining process.

The traditional statistical methods are commonly used in marketing research but our aim is to apply modern approaches of artificial intelligence tools on data for the marketing research which deals with consumer behavior in the market.

The problem of customer behavior falls into the field of marketing. The issues of customer behavior falls into two category, the recognition & understanding of how customers think, feel evaluate, choose among different alternatives, how customers are influenced by their surroundings, how they act during the decision making and purchasing how is their behavior limited by their knowledge are ability to process information, what motivates them and how they differ in their decision making in different ways depending on the importance or product interest (Solomon, 2004).

For the purposes of marketing research two types of data are used firstly, primary data which are obtained from survey (Turcinkova, 2007) and secondary from national and international sources such as statistical offices, states of countries etc. The majority of data that are provided in statistical offices today exist in electronic form. Even data from the surveys are usually recorded electronically or they are converted into electronic form.

Prediction, marketers can surprise their customers and make the customers shopping express becomes a pleasant one.

Yu Minchiang *et al.* (2005) has constructed a marketing decision model which utilized the demographic and geographic variables as input of three individual classifier's-BP networks, decision tree and mahalanobis distance to predict a new customers value.

According to N. Takagi (2006) decision trees and binary decision rules are key techniques in Data mining and knowledge discovery in databases.

Hemant Kumar *et al.* (2011) have applied Data mining in marketing by building a model of the real world based on data selected from a variety of sources which may include corporate transactions, costumer histories and demographic information, process control data and relevant external databases.

Ruxandra Petere (2013) has presented that the main features of a Data mining solutions which can be applied for the business environment and the architecture.

Data mining techniques have been used by Radhakrishnan *et al.* (2013) to uncover hidden patterns & predict future trends and behaviors in financial market.

Khan *et al.* (2013) have used decision tree algorithm for gender classification of frontal images due to its distinctive features according to them their technique demonstrates robustness and relative scale in various for gender classification.

Lilian Sing'oei *et al.* (2013) have presented in their research work Data mining frame work for direct marketing in banking sector, they have provided a comprehensive frame work to guide research efforts focusing on direct marketing strategy and aid practitioners in their guest to achieve direct marketing success using Data mining methods.

Pradyna Mulley *et al.* (2015) have used Data mining technique for customer segmentation in on line retail industry.

7.1.2 Types of Marketing

Generally, there are two types of marketing approaches

7.1.2.1 Mass Marketing

Mass marketing uses mass media such as television, Radio & News papers, and broadcast messages to the public without discrimination. It used to be an effective way of promotion when the products were in great demand by the public.

However, in today's world where products are overwhelming and market is highly competitive,

Mass marketing has become less effective. The response rate, the % of the people who actually buy the products after seeing the promotion is often very low.

7.1.2.2 Direct Marketing

The second approach of promotion is direct marketing. Instead of promoting to costumers in discriminately, direct marketing studies customer's characteristics and needs and selects certain costumers as the target for promotion. The hope is that the response rate for the selected customers can be much improved. In the present chapter Data mining has been applied in the marketing domain.

7.1.3 Important Marketing Areas

- **Banking industry**

Data mining techniques have a no of applications in the banking industry (Vikas Jayashree, 2013) include in the following:

(i) **Credit scoring:** Data mining technologies distinguish the factors like customer payment history which can have higher or lower influences over loan payment.

(ii) **Predict customer profitability:** It identifies pattern based on various factors like products used by a customer in order to predict the profitability of the customer.

(iii) **Customer retention:** The application of Data mining technique identify customer shopping pattern and adjust the product portfolio pricing options provided.

(iv) **Customer Segmentation:** It establishes customer groups and includes each new customer in the proper group.

- **Insurance Industry**

Data mining techniques have many applications in the insurance marketing,can improve it by analyzing the amounts of data available for companies the Data mining techniques have also following applications:

(i) **Summer segmentation ad retention:** Data mining technique establishes customer groups and includes each new customer to the appropriate group identify counts and packages which world increase customer loyalty.

(ii) **Fraud Detection:** Establish patterns of fraud and analyzed the factors which indicates a high probability of fraud for a claim.

(iii) **Risk factor identification:** Analyze the factors like customer claims history are behavior patterns that can have a stronger or weaker influence over earns that can have a stronger or weaker influence over the insurers' level of risk.

- **Retail Industries**

In techniques have several applications in the retail industry (Jiawe et al) including the following:

(i) Establish customer shopping behavior

Data techniques in retail industries identify customer buying patterns and determine what products the unformed is likely to bay next.

(ii) Customer segmentation

Identify customer group associate each customer to the proper group.

(iii) Customer retention

Identify customer shopping patterns and adjust the product portfolio, the pricing and the promotion officered.

(iv) Analyze sales campaigns

Product the effectiveness of a sales campaign the certain factors like the discounts offered for the advertisement us.

7.1.4 Data mining tools

Various Data mining tools (Ruxandra Petere,2013) commercially available for performing advanced data analysis on large volumes of data are shown below in table, various Data mining techniques are used to implement these tools.

Table 2:

Category	**Data mining tools**	**Main features**
Primary tools	Statistical data miner	Statistics and analytics software package which provides data analysis. Data management, statistics, Data mining and data visualization functions
		Provides effective data preprocessing cleaning, and filtering tools, along with tools for producing prediction models in various formats.
		Data mining methods available include; classification, regression, clustering association and sequence analysis
		Applications domains in which it can be used are customer segmentation. Customer retention, credit scoring market basket analysis or price optimization
	IBM SPSS Models	Data mining and text analysis software application used for building predictive models
		Intuitive graphic user interface which allows users to import, manage and analyze their data.
		Data mining techniques included are clustering (k means, support vector machine),classification (Bayesian networks regression, neural networks, decision trees) association rules(Apriori) anomaly detection application domains for which it can be used forecasting sales, customer relationship management, risk management or fraud detection.

	SAS Enterprise Mier	Software applications which provides Data mining algorithms for creating predictive and descriptive models.
		Comprises an early to use graphics user interface which helps with data preparations, summarization and exploration as well as advanced predictive and descriptive modeling.
		Data mining techniques applied include classification (decision trees, neural networks) clustering regression, association rules.
		Tasks or which it can be used: detect fraud, anticipate resource demands, increase acquisitions and curb customer dietician.
Secondary Tools	Microsoft SQL Server Analysis Services	OLAP, Data mining and reporting tools in Microsoft SQL server
		Used to create, manage and explore Data mining models, and then create predictions by using those models.
		Data mining algorithms types included are: Classification, regression clustering, association algorithms, and sequence analysis.
		Application domains for which it can be used: Customer segmentation, forecasting sales, market basket analysis, identifying customers shopping behavior.
Frequently used tools	Oracle Data mining	Embeds Data mining techniques within the oracle database.
		Provides means for building testing, validating, management and deploying Data mining models inside the data base environment
		Supports the following Data mining functions: Classification, regression attribute importance, anomaly detection, clustering, association models and feature extraction
		Application domains for which it can be used: Customers segmentation, recommit next likely product, credit scoring customers profitability or fraud detection

7.1.5 Construction of Decision Tree for Marketing

Fig-14 describes a simple example of construction of decision tree which indicates whether or not a potential customer will respond to a direct mailing. Here internal nodes are represented as squares whereas leaves are denoted as ellipse. Two or more branches may grow form each internal node (i.e. not a leaf). Each node corresponds with a certain characteristic and the branches correspond with a range of rules. These ranges of values must give a partition of the set of values of the given characteristics.

The above decision tree incorporates both nominal and numeric attributes. Even this classifier, the analyst can predict the response of a potential customer (by sorting it down the tree) and understand the behavioral characteristics of the entire potential customer population regarding direct mailing. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values. In case of numeric attributes, decision trees can be geometrically interpreted as a collection of hyper planes, each orthogonal to one of the axes.

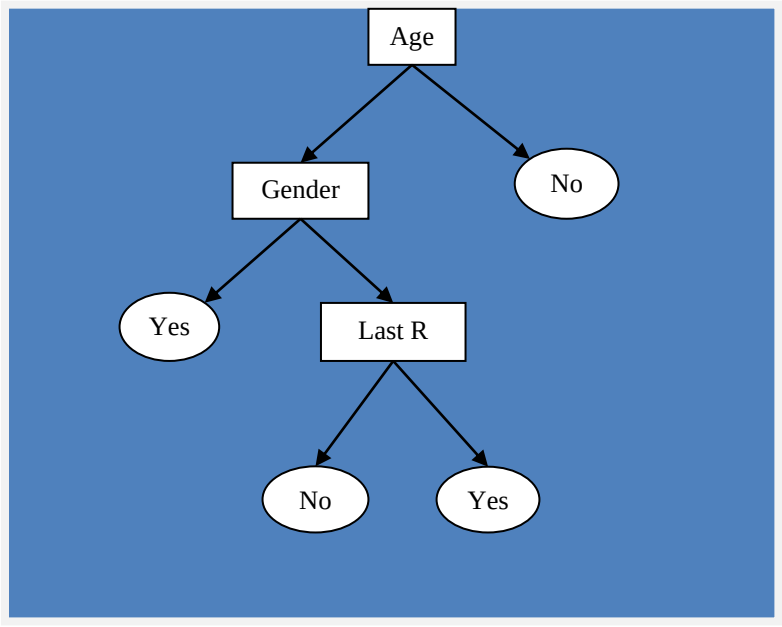


Figure 14: Decision tree showing response to direct mailing

7.1.6 Experimental Analysis

Table 3:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	500
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	25000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	12500
(iv)	Total Cost of Promotion	25,00,00	37500
(v)	Response rate	1%	3%
(vi)	No of Sales	50	15
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	67500
(viii)	Net Gain from Promotion	2,50,00	30000

Table 4:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	1000
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	50000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	25000
(iv)	Total Cost of Promotion	25,00,00	75000
(v)	Response rate	1%	3%
(vi)	No of Sales	50	30
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	135000
(viii)	Net Gain from Promotion	2,50,00	60000

Table 5:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	1500
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	75000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	37500
(iv)	Total Cost of Promotion	25,00,00	112500
(v)	Response rate	1%	3%
(vi)	No of Sales	50	45
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	202500
(viii)	Net Gain from Promotion	2,50,00	90000

Table 6:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	2000
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	100000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	50000
(iv)	Total Cost of Promotion	25,00,00	150000
(v)	Response rate	1%	2.75%
(vi)	No of Sales	50	55
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	247500
(viii)	Net Gain from Promotion	2,50,00	97500

Table 7:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	2500
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	125000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	62500
(iv)	Total Cost of Promotion	25,00,00	187500
(v)	Response rate	1%	2.50%
(vi)	No of Sales	50	62.5
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	281250
(viii)	Net Gain from Promotion	2,50,00	93750

Table 8:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	3000
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	150000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	75000
(iv)	Total Cost of Promotion	25,00,00	225000
(v)	Response rate	1%	2.25%
(vi)	No of Sales	50	67.5
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	303750
(viii)	Net Gain from Promotion	2,50,00	78750

Table 9:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	3500
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	175000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	87500
(iv)	Total Cost of Promotion	25,00,00	262500
(v)	Response rate	1%	2%
(vi)	No of Sales	50	70
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	315000
(viii)	Net Gain from Promotion	2,50,00	52500

Table 10:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	4000
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	200000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	100000
(iv)	Total Cost of Promotion	25,00,00	300000
(v)	Response rate	1%	1.75%
(vi)	No of Sales	50	70
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	315000
(viii)	Net Gain from Promotion	2,50,00	15000

Table 11:

S. No	Details	Mass Mailing	Direct Mailing
(i)	No of Customers mailed	5,000	4500
(ii)	Printing & Mailing Cost **(Rs,50.00 each)**	25,00,00	225000
(iii)	Data mining Cost **(Rs 25.00 each)**	Nil	112500
(iv)	Total Cost of Promotion	25,00,00	337500
(v)	Response rate	1%	1.70%
(vi)	No of Sales	50	76.5
(vii)	Gain from Sale **(Rs.4500.00 each)**	22,50,00	344250
(viii)	Net Gain from Promotion	2,50,00	6750

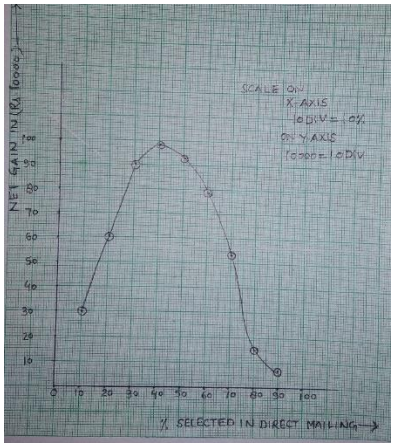


Figure 15: Direct mailing graph by experimental analysis

7.1.7 Results & Discussions

From the curve shown in Fig-15, it is evident that if we mail only to some top small percent of customers then the total number of responders is too small and net profit is also small. If justifiable for the printing and mailing cost, hence there is an optimal percent which can bring us the maximum net profit. Such optimal cut off points depend critically on many factors in the whole process of direct marketing using Data mining; cost of mailing, cost of Data mining profit presale etc.

Some of these factors depend on each other, making the strategic decision on Data mining non trivial, here we demon stated that Data mining is an effective tool for direct marketing which can bring more profit tool retail industry banks and insurance companies than the traditional means of mass marketing.

Such mythology would be effective and easily transferable to market managers and also helps to visualize knowledge underlying the data set. The architecture proposed for the Data mining solutions for the business environment would improve the efficiency of a company by providing valuable discussion making knowledge to minimize operating costs and gain competitive advantage.

Our current society needs Data mining for improving many domains of human life. Marketing areas like retail, banking and insurance can use Data mining methods to improve customer experiences, make optimal decision, strength their market position and achieve competitive advantage. There are various commercially available Data mining tools to provide support for fulfilling these requirements.

7.2 Data Mining in Healthcare

7.2.1 Introduction

Data mining has been used intensively & extensively in healthcare organizations because of following reasons.

Firstly, the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining is a powerful tool which provides the methodology and technology to transform these mounds of data into useful information's for decision making.

Secondly, the existence of medical insurance fraud and abuse which have lead many healthcare insurers to attempt to reduce their loses by using Data mining tools to help them find and track offenders (Christy,1997).Recently there has been reports of successful Data mining applications in healthcare "fraud and abuse" detection (Milley,2000).

Thirdly, Data mining applications can benefit healthcare providers such as hospitals, clinics, and physicians and patients by identifying effective treatments and best practices. (Gillespie, 2000; Kolar, 2001)

Fourthly, the use of Data mining applications in healthcare is the realization that Data mining can generate information which are very useful to all parties involved in the healthcare industry.

This chapter explores Data mining applications in healthcare. It discusses Data mining up to data healthcare evidence and improves documentation. Moreover, the efficiency of the healthcare delivery is improved by reducing costs through faster order processing or eliminated destination of text.

Diabeties: Diabeties mellitus which is frequently called as diabeties is a situation produced by a decrese fabrication of insulin and due to this glusoce levels in the blood is going to be high.This happens because either the pancreas is unable to produce enough insulin or the cells in the body have become resistant to insulin.Diabeties affect the capability of human body to use the energy present in food material.

The types of diabeties are described below.

Type-1: In this type of diabeties pancreas does not produced adequate amount of insulin as a result of which the level of glusoce in blood increases from typical range.The persons suffering from this type of diabbeties are usually dependent on external insulin injected in the body after regular intervals.It is caused by a gentic prediscriptions,medical dangerous associated with this

type of diabetic include diabetic rationpathy(Eyes disorders),diabetic neuropathy(nerves disorder) and diabetic neuropathy(kidneys disorder).It counts for 95% diabetic cases.

Type-2: In this type of diabetes the body is not able to consume the insulins properly.Due to insulin resistance,this is usually generated due to obesity and overweight children. This is non insulin dependent and milder than type1 diabetes. It generates major effects on heart deases and geart strokes.It is not cureable but can not be controlled with proper neutrions, exercise and weight management.

Gestational diabetes: In this type of diabeties the married woman who are not affected with diabetics according to previous medical history but high glucose level is diagnosed during/after pregency. According to national institute of health the reported rate of jusitional diabeties is between 2% to 10% of progenies.

Hian Chye kob *et al.* (2011) has explored in their article, the Data mining applications in healthcare domain. They have discussed Data mining application within healthcare in major area like evaluation of treatment effectiveness management of healthcare customer relationship management and the detection of fraud and abuse.

Abdullah *et al.* (2013) have shown in their article the predictive analysis of diabetic treatment using a regression-based Data mining technique. The oracle data miner (ODM) was employed as a software mining tool for predicting modes of tracking diabetes. The support vector machine algorithm was used for experimental analysis.

Monali *et al.* (2014) have mentioned in their article about analysis of the uniqueness of medical Data mining, overview of healthcare decision support systems currently used in medicine, identification and selection of the most common Data mining algorithms implemented in the modern HDHSS and Companion between different algorithms in Data mining.

Currently the rate of data accumulation is much faster than the rate of data interpretation. These data need to be effectively organized and analyzed in order to be useful.

7.2.2 Data Mining Applications in Some Healthcare Arena

Successful Data mining applications have been implemented in some healthcare arena as described below:

- **Healthcare Management**

To enhance healthcare management Data mining applications can be developed to better identify and track chronic disease states and high risk patients, design suitable interventions and reduce the number of hospital admission and claims.

Data mining initiates to reform outputs and minimize examples through better diseases management. For instance it uses emergency cell and hospitalization claims data, pharmaceutical records and physician interviews to identify unknown asthmatics and develop appropriate interventions (Kincade, 1998).Data mining can also be used to identify and understand high cost patients (Silver, 2001).Data mining can be used to analyze massive volume of data and statistics to search for patterns that might indicate an attack on bioterrorists (Piazza, 2002).

The Light Weight Epidemiological Advanced Detection Emergency Response System (LEADERS) is one such efforts.In the past, LEADERS have uncovered several disease outbreaks.

- **Hospital Infection Control**

Data mining can also be used for hospital infection control (Kruze, 2001) or as an automated early warning system in the event of epidemics. A syndromic system based on patterns of symptoms is likely to be more efficient and effective than a traditional system that is based on diagnosis. An early warning of the global spread of the SARS virus is an example of the usefulness of a syndromic system based on Data mining.

Computer assisted surveillance research has focused on identifying high risk patients, expert systems and possible cases and detecting deviations in the occurrence of predefined events. The system uses associations rules on culture and patient care data obtained from the laboratory information management systems and generates monthly patterns that are reviewed by an export in infection control. Developers of the system conclude enhancing infection control with the Data mining system is more sensitive than traditional infection control surveillance and significantly more specific.

- **Comparison across Healthcare Groups**

At a higher level Data mining can facilitates comparisons across healthcare groups of things such as practice patterns, resource utilization, and length of stay and costs of different hospitals. (Johnson, 2001).

Recently an Urban country (Sierra) has used Data mining extensively to identify areas for quality improvements including treatment guidelines disease management groups and cost management (Scheurberg, 2003)

- **High Risk Patients identifications**

To identify high risk patients the predictive modeling technology of Data mining is used. Enormous information of patients is combined and explored to predict the likelihood of short term health problems and intervene proactively for better short term and long term results.

A robust Data mining and model building solution identifies patients who are trending toward a high risk condition. This information gives nurse care coordinators steps can be taken to improve the patient's quality of healthcare and to pervert health problems in the future.

- **Effectiveness of Medical Treatments**

By comparing and contesting causes, symptoms and courses of treatments Data mining can deliver an analysis of which courses of action prove effective (Milley,2000).For example the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compound to determine which treatments work best and are most cost effective (Kincade, 1998).

Data mining can help identify successful standardized treatments for specific diseases. In 1999 Florida hospital launched the clinical best practices initiative with the goal of developing a standard path of core across all campuses, clinics and patient admissions. (Kolar, 2001)
Limited healthcare has mined its treatment resend data to explore ways to cut costs and deliver better medicine (Young,1997).It has also developed clinical profiles to give physicians information about their practice patterns and to compare these with those of other physicians and poor reviewed industry standards.

Other Data mining applications related to treatments include associating the various side effects of treatments, collating common symptoms to and diagnosis determining the most effective drug compounds for treating sub populations that respond differently from the mainstream populations to certain drugs and determining proactive steps that can reduce the risk of affliction (Milley,2000).

- **Fraud and Abuse**

Applications of Data mining which attempt to find fraud and abuse after establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, laboratories or others. These applications, among other things can high light inappropriate prescriptions or referrals and fraudulent insurance and medical claims.

7.2.3 Healthcare Decision Support Systems

- **Help**

One of the most popular and advanced healthcare decision support system is called Help. It helps the clinicians is interpreting healthcare information, diagnosing the diseases of patients, maintaining healthcare protocols and other tasks.

- **DX plain**

It is a healthcare decision support system (HDSS) available through the World Wide Web which assists clinicians by generating stratified diagnoses based on user input of patient signs and symptoms, laboratory results, and other healthcare findings. Each healthcare finding interred into DX plain is assessed by determining the importance of the finding and how strongly the finding supports a given diagnosis for each disease in the knowledge base. Using this criterion, DX plain generates ranked differential diagnoses with the most likely diseases yielding the lowest rank.

- **ERA (Early Referrals Application)**

It is one of the newest and most promising Healthcare Decision Support Systems. This solution is dedicated to detection of different types of the cancers in their early stage. The Doctor then takes the output of the DDSS and point out which are relevant and which are not. Another important classification of a HDSS is based on the timing of its use.

Doctor's use these systems at point of care to help them as they are dealing with a patient with the timing of use as either pre diagnose, during diagnoses, or post diagnoses. Pre- diagnoses HDSS systems are used to help the physician prepare the diagnosis. HDSS used during diagnosis help review and filter the physician's preliminary diagnostic choices to improve their final results.

And post-diagnoses HDSS systems are used to mine data to derive connections between patients and their past medical history and to predict future events.

7.2.4 Characteristics of Healthcare Decision Support Systems

The healthcare DSS's are the type of computer programs which help physicians and medical staff in healthcare decision making tasks (Abbasi *et al.* 2006). Following are the characteristics of healthcare decision support systems.

(i) Most of the healthcare decision support systems are equipped with diagnostic assistance module, therapy critiquing and planning module, medications prescribing module, information retrieval subsystem (for instance formulating accurate clinical questions) and image recognition and interpretation section, (X-Rays, CT, MRI scans). Interesting examples of HDSS's are machine learning systems that are capable of creating new healthcare knowledge.

(ii) By analyzing healthcare cases a Healthcare Decision Support System can produce detailed description of input features with a unique characteristics of healthcare conditions.

(iii) It supports may be priceless in looking for changes in patients health condition. Theses systems may improve patient's safety by reducing errors in diagnosing. They may also improve medications and test ordering.

(iv) More over the quality of care gets better due to the lengthening of the time clinicians spend with a patient. It may bean effect of applications of proper guidelines.

7.2.5 Limitations

Data mining applications greatly benefit the healthcare industry. However there are certain limitations as described below:

Healthcare Data mining can be limited by the accessibility of data, because the raw inputs for Data mining often exist in various settings and systems such as administration, clinics, laboratories and more. Hence data have to be collected and integrated before Data mining can be applied while several authors and researchers have suggested that a data warehouse to be built before Data mining is attempt that can be costly and time consuming project. A data warehouses is successfully built by in intermountain healthcare from five different sources a clinical data repository, auto care case mix system, labor arty information system, ambulatory case mix system and health proems data base and used to find and implement better evidence based clinical solutions.A distributed network topology interested of a data warehouse has been suggested by Ookley (1999) for more efficient Data mining.

Secondly,other data problems may arise these include missing, corrupt in consistent or non standardized data, such as pieces of information recorded in different formats in different data sources.

In particular the lack of standard chemical vocabulary is a serious hindrance to Data mining, cios and more (2002) have argued that data problems in healthcare are the result of the volume, complexity and heterogeneity of medical data and their poor mathematical characterization and non canonical form. Further these may be legal, ethical and social issues such as data ownership and privacy issues related to healthcare data. Thus the quality of Data mining results and applications depends on the quality of data (Choprian *et al.* 2001)

Thirdly a sufficiently exhaustive mining of data will certainly yield patterns of some kind that are a protect of random fluctuations. This is especially true for large data set with many variables.

Concerns in their cavly stage. The applications have been developed in Great Britain by GP's associated with the university hospitals of Leicester NHS Trust Since 2001.

Hence many interesting or significant patterns and relationships focus in Data mining may not be careful. It has been warned by Murray (1997) and Hand (1998) against using Data mining for data dredging or fishing which is randomly traveling through data in the hope of identifying patterns.

7.2.6 Construction of Decision Tree for Medical Application

Let us imagine that we want to develop a medical system for diagnosing patients according to the results of several medical tests. Based on the result of one test, the physician can perform or order additional laboratory tests. Fig-15 shows the diagnosis process, using decision trees of patients which suffer from a certain respiratory problem.

The decision tree employs the following attributes: CT finding (CTF); X-Ray finding (XRF); Chest Pain Type (CPT); and blood test finding (BTF). The physician will order on X–Ray, if chest pain type is "1" however, if chest pain type is "2", then the physician will not order an X-ray but will order a blood test. Thus medical tests are performed just when needed and the total cost of medical tests is reduced.

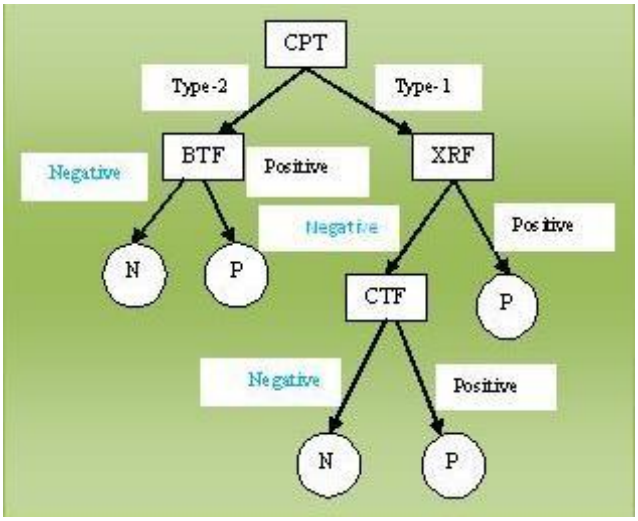


Figure 16: Decision tree for Medical Application

7.2.7 Experimental Analysis

To discuss a data mining application in healthcare we take Healthcare data from various health care institutions (Hospitals, Clinics, and dispensaries etc.) situated in Tarai region (Gorakhpur, basti, Devoria, Siddharthnagar, Kushinagar, etc. Districts) with the help of questionnaire. The No of patients examined was 2090 (including diabetic patients).

Here we are interested to find out how certain variables are associated with the onset of diabetes. Thus the aim of this Data mining application is to identify the risk factors associated

with the onset of diabetes so that appropriate informations can be communicated to patients. The data set contains nine variables of interest: Gender, Age, Body Mass Index (BMI), Waist Hip Ratio (WHR), Height, Smoking status Religion, Cast, & the number of times a patient exercises per weak.

The data set comprises of 282 or 13.49% of positive diabetic cases and 1808 or 86.50% of negative non diabetic cases.

After reviewing the work of Breault *et al.* (2002) on the data mining of a diabetic data source we have decided that decision trees are an appropriate data mining technique to use to find out how certain variables are associated with the onset of diabetics.

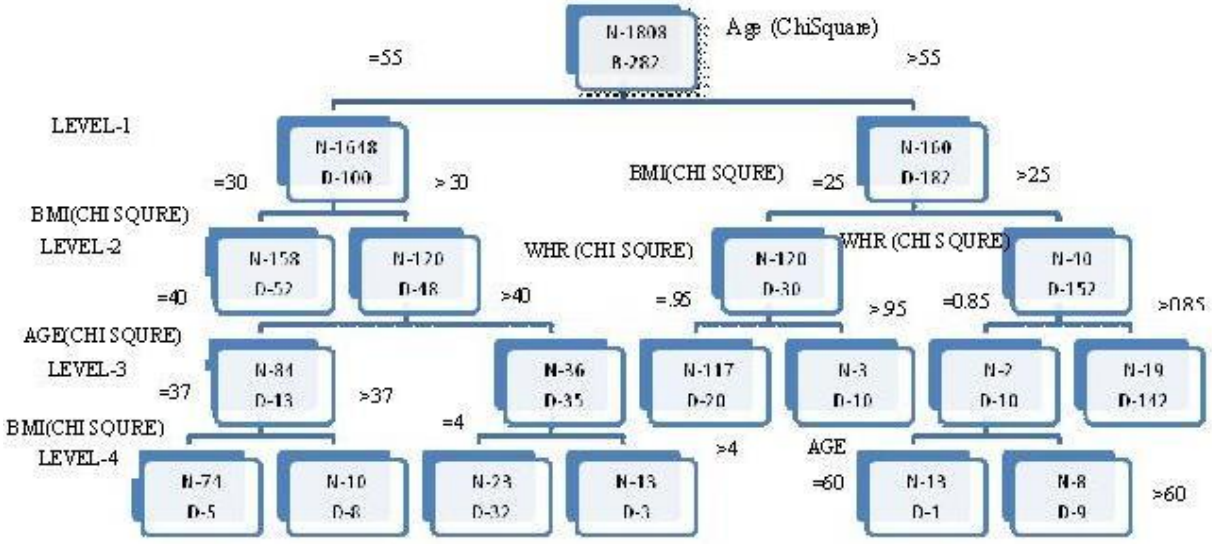


Figure 17: Visualization of the decision tree results and facilitates interpretation

7.2.8 Results and Discussions

Results are summarized as shown in the Fig-16. Age, Body Mass Index (BMI), Waist Hip Ratio (WHR) and the No of exercise per week are significantly associated with the onset of diabetics. Fig-17 gives a visualization of decision tree and facilitates interpretation of the results. The generated decision tree can be interpreted as follows. The result shows that the age is the most important factor associated with the onset of diabetes (level 1 Fig-17) with individuals older than age 55 showing significant higher risk of diabetes, compared with their counterparts.

At the next level 2 (Fig-17) the Body Mass Index (BMI) is the next most important factor associated with the onset of diabetes. In particular individuals younger than 30 with a body mass index (BMI) of less than 28.57 have a very low risk of diabetes (probability of only 3.29% in the

cohort). Moreover, increasing levels of BMI is associated with increasing risk of diabetes. For individuals older than 55 with a BMI greater than 25, the risk is 79.16%.

The Waist Hip Ratio (WHR) is the next most important factor (level 3 Fig-17) with increasing WHR associated with an increased risk of diabetes. For example, the highest risk individuals in the data base are those above 55 years old with BMI of more than 25 and WHR above 0.85 –their risk probability is 88.12 in this cohort.

The exercise per week (level 4 Fig-17) is the next most important factor associated with the onset of diabetes. The greater the no of exercise per week the lower is the no of diabetic cases.

In a similar manner the remaining nodes in the decision tree can be interpreted. Nodes further down in the decision tree are less important in view of their smaller sample sizes and also because of their more restrictive sub-setting at lower levels. It is also concluded that gender & religion is not effective attributes associated with the onset of diabetes. Thus decision tree can help health organizations to identify the high risk individuals and appropriate messages can be communicated to them. For example, healthcare organizations can launch a health promotions campaign to educate people that large BMI and WHR are risk factors associated with the onset of diabetes. It can also scan through its patient data bases to identify individuals for further counseling or medical checkups.

7.2.9 Future Direction

In healthcare Data mining applications can have extreme potential and usefulness. However the success of Data mining in healthcare bringes on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry consider how data can be better collected, stored, prepared and mined.

Possible directions of Data mining in healthcare include the standardizations of clinical vocabulary and the sharing of data across organizations to enhance the benefits of health care Data mining applications. Further as healthcare data are not limited to just quantative data, such as physicians notes or clinical records, it is necessary to also explore the use of text mining to expand the scope and nature of what healthcare Data mining can correctly do. In particular it is useful to be able to integrate data and text mining.

Finally it is hoped that this research work make a contribution to the Data mining and healthcare literature and practice. It is also hoped that this work can help all parties involved in healthcare reap the benefits of healthcare Data mining.

7.3 Data Mining In Terror Related Activities

7.3.1 Introduction

In the wake of terrorist attacks on America on 11 Sept 2001, England on 07 July 2005, Russia on 24 Jan. 2011, France on 13-14 Nov. 2015 & India 26-29 Nov. 2008 etc. most of the countries became more vigilant about their national security. In the light of these, several techniques have been applied to assist the law enforcement agencies to identify terrorists and to prevent terrorism. One of such techniques is the use of computer technology and computer analysis for effective analysis of terrorist activities. Data mining can be used by law enforcers to analyze information's by applying several Data mining techniques. In this chapter we have discussed how Data mining techniques can be adopted by law enforcement agencies in tracking the activities of terrorists and their criminal activities. The limitations of Data mining against terrorism have also been discussed in the present chapter.

Sacham *et al.* (2012) have proposed a TGP Model to predict the terrorist group in India using the historical data. The database has been taken from GTD that includes the terrorist attacks in India from 1998 to 2008.The researchers have used the terrorist corpus, parameters weight and value as input. The unsupervised learning clustering technique is applied to form the clusters of the data.

The mathematical equation is also used to perform some main steps. The overall performance aimed by the proposed model is 80.41%.

Elovici *et al.* has presented in his research paper an innovative known ledge based methodology for terrorist detection by using web traffic contents. Their proposed methodology learns the typical behavior (profile) of terrorists by applying a Data mining algorithm to the textual content of terror related website.

Chaurisia *et al.* (2012) have used Data mining technique as one of the effective solution in social network analysis which studies terrorists networks for the identification of relationship and association that may exist between terrorist nodes. According to them terrorist activities can also be detected by means of analyzing web traffic content.

The authors G. Faryral & B. H. Wasi (2014) have proposed novel ensemble framework for the classification and prediction of terrorist group in Pakistan which consists of four base classifications namely, NB, K-NN, ID3 and Decision stumps. Majority vote based ensemble technique is used to combine these classifiers. It has been concluded that the new approach

achieves a considerably better level of accuracy and less classification error rate as compared to the individual classifier.

Vighne *et al.* (2016) have developed a way to detect terrorist activity on the web by eavesdropping on all traffic of websites associated with terrorist groups organizations in order to detect the accessing users based on their IP address because terrorists are using different IP address changing them frequently such that it will become more hard to identify them and to detect them.

7.3.2 Real -Time Threats

In the case of real time threats there are timing constraints. That is, such threats may occur within a certain time and hence we require to respond to it immediately. Ex. of such threats are the spread of smallpox virus, chemical attacks, nuclear attacks, network intrusions, bombing of a building before 10 am in the morning etc.

7.3.3 Non Real -Time Threats

Essentially with non real time threats, we have time together data, build say profiles of terrorists, analyze the data and take actions. Now, a non real time threat could become a real time threat. That is, the Data mining tool could state that there could be some potential terrorist attacks. But after a while, with some more information, the tool could state that the attacks will occur between Sept 10, 2001 and Sept 12, 2001.Then it becomes a real time threat. The challenge will then be to find exactly what the attack will be?

Will it be an attack on the world trade Center or will it be an attack on the Tower of London or will it be an attack on the Eiffel tower? We need Data mining tools that can continue with the reasoning as new information comes in. That is as new information comes in the warehouse needs to get updated and the mining tools should be dynamic and take the new data and information into consideration in the mining process.

Non real time threats are threats which do not have to be handled in real time, i.e. there are no timing constraints for these threats. For example we may need to collect data over months, analyze the data and then detect and/or prevent some terrorist attack which may or may not occur. The question is how does Data mining help towards such threats and attacks? As we have stressed in we need good data to carry out Data mining and obtain useful results. We also need to reason with in complete data. This is the big challenge as organizations are often not prepared to share the data. This means that the Data mining tools have to make assumptions about the data belonging to other organizations.

The other alternative is to carryout federated Data mining under some federated administrator. For example the homeland security insures that the various agencies have autonomy but at the same time collaborate when needed.

Next we need to start gathering information about as many people as possible because some times even those who seem most innocent may have ulterior motives, one possibility is to group the individuals depending on say where they come from, what they are doing, who their relatives are etc. Some people may have more suspicious backgrounds than others. It we know that someone has had a criminal record then we need to be vigilant about that person.

Again to have complete information about people, we need tighter all kinds of information about them. This information can include information about their behavior, where they have lived, their religion and ethnic origin. Their relatives and associates, their travel record etc.

Gathering such information is a violation to one's privacy and civil liberties. The question is what alternative do we have? By omitting information we may not have the complete data not only about individuals but also about various events and entities.

Once the data is collected the data has to be formatted and organized. Essentially one may need to build a warehouse to analyze the data. Data may be structured or unstructured data. Also there will be some data which is warehoused and may not be of much use.

Once the data is gathered is organized the next step is to carry out mining. Then comes the very hard part, how do we know that the mining results are useful? There can be false positives and false negatives. A false positive is when a process incorrectly reports that it has found what it is looking for. A false negative is when it incorrectly reports that it has not found what it is looking for.

7.3.4 Types of Disasters

These are mainly four types of disasters as mentioned below:

7.3.4.1 Natural Disasters

By natural disasters we mean disasters due to hurricanes, Earthquake, Fires, Power failures and accidents. Some of these disasters may be due to human errors such as pressing the wrong switch in a process plant causing the plant to explode. Data mining can help to detect some of the natural disasters, i.e. by analyzing a lot of geological data, the Data mining tool may predict that an earthquake is about to occur in which case the people in the area could be

evacuated beforehand. Similarly by analyzing the weather data the tool could predict that Hurricanes are about to occur.

The long term measures to be taken for natural disasters may be quite different from terrorist attacks. It is not everyday that we have an earth quake, even in the most earthquake prone regions. Therefore we have time to plan and react. This does not mean that a natural disaster is less complex to manage. It could be devastating and take many humans lives. However countries usually plan for such disasters mainly through experiences.

7.3.4.2 Disasters due to human errors

Disasters due to human errors are also a source of major concern. We need to continually train the operators and give them advice to be cautious and alert. We need to take proper actions if humans have been careless, i.e. unless there is an absolutely good excuse, human errors should not be treated lightly. This way human will be cautious and perhaps not make such errors.

7.3.4.3 Disasters due to terrorists attack

Disasters due to terrorist attacks are quite different. The problem is, one does not know when it will happen and how it will happen. Many of us could never have imagined that airplanes would be used as weapons of mass destruction to bring the famous world trade center towers down. Many of us still may not know what the next attack by chemical weapons or by cyber terrorism. The counter measures for prevention and detection may be quite intense for terrorists attack.

Our goal is to examine the various Data mining techniques to see how they could be applied to handle the various threats created by terrorists that have been discussed, mentioned ahead.

However it should be noted that to develop effective technique the Data mining specialists have to work together with counter terrorism experts. That is one can not use the techniques without a good understanding of what the threats are. Therefore while the contents of this thesis may be used as a reference i.e. would like to urge those interested in applying Data mining techniques to solve real world problems and terrorists attacks to work with counter terrorism specialist.

7.3.4.4 Attacks by Malicious Intrusions

In malicious intrusions, intruders try to tap into the networks & get the information i.e. being transmitted. These intruders may be human intruders are Trojan horses setup by humans. Intrusions could also happen on files. These intrusions could be intruding the networks, the web clients and servers, the databases, operating systems etc. For example one can masquerade as

some one else and log into someone else's computer system and asses the files. Intrusion can also occur on databases. Intruder poising as legitimate users can pose queries such as SQL queries and access the data they are not authorized to know.

7.3.5 Types of Terrorist Attacks

There are various types of terrorist attack as under:

7.3.5.1 Chemical attack

Attacks using chemical weapons by the terrorists are extremely dangerous. Terrorists can spray poisonous gas and other chemical materials in to air, water and food supplies. For example various dangerous chemical agents can be spread from the air on plants and crops. These plants and crops can get into the food supply and millions of human kinds, animals and birds etc.

We have to develop technologies like Data mining to detect and prevent such deadly attacks. There are some excellent references about such terrorist's activities.

7.3.5.2 Nuclear attacks

Another form of deadly attacks by the terrorists is the nuclear attacks. Such attacks can wipe out the entire population in the world. There are various nations developing nuclear weapons when although these nations do not have the authorizing to develop illegally. This is what that makes the world vary dangers we have to generate Data mining technologies to detect and prevent such dangerous attacks.

7.3.5.3 Bio terrorism

Such attacks can kill several million people, animals, birds' insects etc, within a very short space of time. Recently there has been an increasing awareness of the dangerous due to bio terrorism tracks resulting in the spread of infectious diseases such as smallpox, yellow fever and similar deceases.

These deceases are so infections that it is critical that there spread is detected as soon as they occur. Our aim is to prevent such attacks in time. One option is to carryout mass vaccination but this would mean some health hazards to various groups of people. Our challenge is to use Data mining technologies to detect and prevent such deadly attacks. Technologies will include censor technologies Data mining and data management technologies.

7.3.5.4 Attacks on critical infrastructure

Attacks on critical infrastructure can destroy a nation and its economy. The infrastructure attacks include attacks on the telecommunication lines, the electronic equipments, power plants

gas reservoirs water supplies units' food supply units and other basic incites which are critical for the operation of a nation.

Attacks on critical infrastructures can occur during any type of terrorist attacks weather they are non information related or information related attacks.

For example terrorists can attack on the software that runs the telecommunication industry and close down all the telecommunication lines. Similarly software which controls the power and gas supplies can be attacked by the terrorists. Such attacks can also occur through bombs and explosives by the terrorists, attacks on transportation lines such as high ways and railways tracks also come as attacks on infrastructure. Our goal is to examine Data mining and related data management technologies to detect and prevent such infrastructure attacks.

7.3.6 Methods and Methodology

7.3.6.1 Classification

The next step is to start searching as to the likely reasons of the attack and the individuals who might have responsible attack, we have already pointed out that terror investigation is the classification:

Wives *et al.* 2000 has pointed out that classification is a Data mining techniques which produces the characteristics to which a population is divided based on the characteristics. According to Thurassingham a classification divides the dataset based on certain fried condition. In case of crime classification assumes that we have some idea of the individuals (suspects) based on the predefined criteria. For example let us suppose that the law enforcement agencies have reported a kidnapping case then they may try to form the idea of the kidnappers, say 3 males between 28 and 32 age, Muslim, speak Urdu frequently between 5' and 7'tall.

Then the classification has been completed and it becomes imperative to place all males meeting the above criteria under suitable observation. The algorithm which is to be used will be such that the population will be divided into two clear parts male and females.

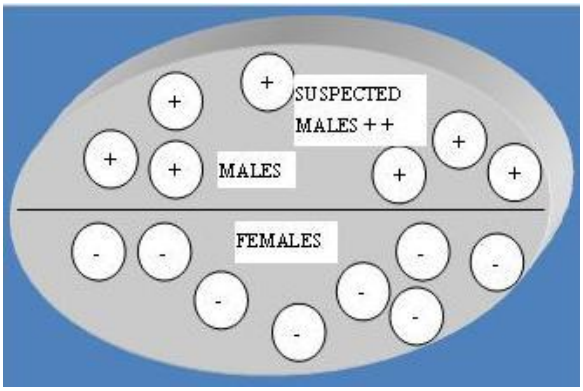


Figure 18: Classification algorithms segmentation of a data set based on some data

The algorithm will also segment the males according to their suspected criteria. Fig-17 indicates a typical segmentation scenario where males are segmented from females. The male data set could be further segmented based on our criteria. As a next step we scan the historical dataset (data ware house) for concerned make her.

If the searches do not produce positive result we could narrow the search by again classifying the classified dataset. Assuming that we were not able to obtain a class the new criteria will then be used to update out historical database (data warehouse) if then another related crime occurs it becomes easier to form a match.

Fig-18 shows a classification algorithm. The data warehouse is classified here based on the criteria for suspects individuals.In our example the data set is classified into males and females hence this classification is not adequate.

Therefore the process continuous until we have a dataset which can be matched against the criteria. Our classified database is now compared against the criteria if we have a related match then it is placed under surveillance otherwise the process continuous until there are no more matches.

If after analyzing a crime with classification technique it is expected that some suspected individuals will come up then to narrow down the suspects, their activities, links, associations, and relations could further be analyzed using the link analysis Data mining techniques.

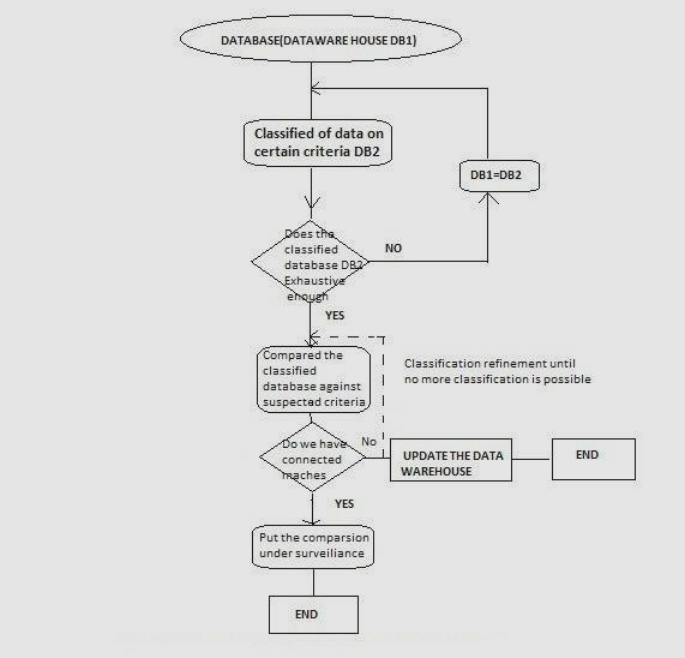


Figure 19: Classification flow for detecting crime & Combating terrorism

7.3.6.2 Link analysis

Link analysis is also a Data mining techniques which is very advantages in extracting valid and useful patterns. The theoretical structure of link analysis is based on the fact that incidents are connected to one another and therefore are mutually exclusive According to link analysis if A is linked to B and B is linked to C and D then A could be indirectly linked to D. If a link analysis structure is visualized, then it turns into the form of graph. This technique uses several graphs.

Memon *et al.* (2005) have pointed out that links analysis; technology can be applied by law enforcers and intelligence analysts to examine graphically the anomalies and inconsistencies, relationships among networks, and contacts hidden in the datasets. of course, la is the first step by means of which networks of people, places organizations, vehicles, bank accounts, mobile cells, email contacts etc can be searched out, linked, assembled, examined analyzed and detected.

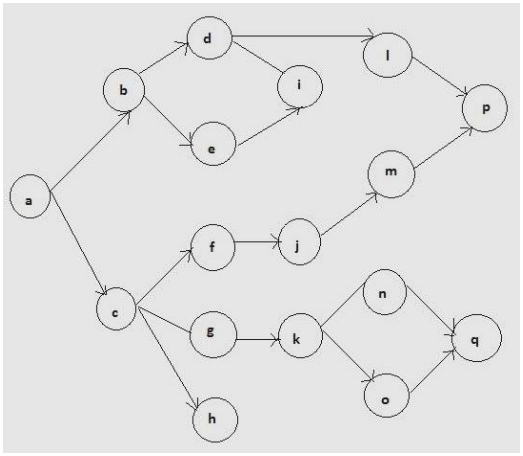


Figure 20: Graph Representation of Link Analysis

The larger is the ability of a representation of link analysis to analyze the graph the larger will be the strength of this technique. Let us consider the above Fig- when the target dataset has been classified we shall have a large number of individuals to part under surveillance we shall have a large number of individuals to put under surveillance. We shall see how we arrive at some suspected individuals we know that terrorists and criminals sell down operate individuals we know that terrorists and terminals sell down operate in a vacuum that is there must be links to other individuals.

Link analysis can be used to analyze the activities of individuals by farming a link of their activities. These connections might be in the form of mobile conversation, places visited, bank transaction etc. The algorithm of link analysis can then make connections between

suspected individuals. We consider the simple graph as shown in Fig-20 which represents the mobile conversation of source suspected individuals, band c can be linked to q while b can be linked to p but same is not true for b to q.

7.3.6.3 Clustering techniques:

According to Han J kamber cluster techniques is the process of portioning data objects (records, documents etc) into meaningful groups or clusters so that objects within a cluster have similar characteristics but are dissimilar to objects in other cluster. Clustering can be understood as unsupervised classification of unlabelled patterns (observations, data items or feature vectors) because no predefined category labels are associated with the objects in the training dataset. Clustering of web documents viewed by internet users can shows collections of documents belonging to the same topic.

Clustering has also been with mankind since very beginning. People cluster tighter according to their certain characteristics, qualities, and attributes people from the same country religion tribe race etc. cluster together.wires et al (2010) have pointed out that data clustering allows us to construct more simpler under stand able modules of that world which can be worked upon more comfortably.

Clustering in another Data mining technique which can because to detect the terrorism and crime. Classification and clustering are almost the same but whereas classification requires basis parameter but clustering does not require any parameter. Clustering techniques and algorithm are dependent on real life model which individual with some virtues must cluster together.

Fig-21 shows a sample of cluster cases of kidnapping in some selected areas. Cluster ssumes that in crime dominates area, individuals with some crime specialties will cluster together.

For example individuals who are specialties in kidnapping tend to cluster together. Clustering technique will then identify given cluster and their areas of operation any time when a crime is reports. Then the law enforcement agencies can observe the related clusters and examine them for clues.

Two crowns has defined that clustering is a way to segregate data into groups which are not previously defined whereas classification is a way to segment data by assigning it to groups that are already defined.

Fig-21 represents a typical clustering algorithm.here the data set is clustered into various clusters, and then crime suspects are placed under each cluster for comparisons. If they tally then they are placed under surveillance a new cluster is then formed and the dataset updated.

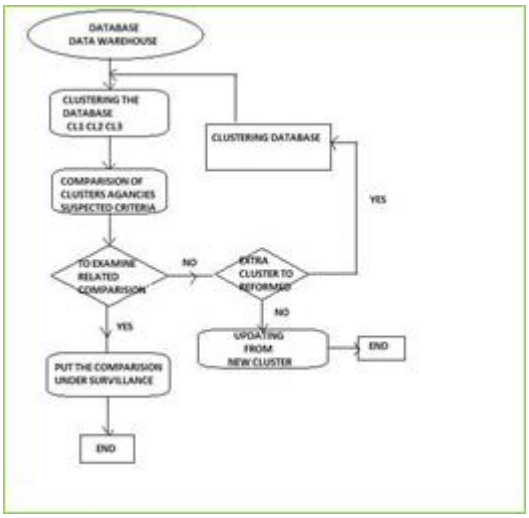


Figure 21: Cluster Algorithm for Crime Identification

Sometimes a clustering anomaly occurs and is referenced anomaly detection. This is the case of a relevant event which happens within a particular place and an event activity happening elsewhere: Fig-21 shows that when an anomaly occurs, it also shows the starting of another cluster. In the Fig l1 and l3 are the normal clusters of kidnapping cases whereas l2, l4, l5 and l6 are all anomalies of kidnapping.

7.3.6.4 K- nearest neighbor method (K-NN):

According to two crowns (2010) K-nearest neighbor (K-NN) is a classification technique that decides in which class to place a new case by examining same number the “K” in K nearest neighbor of the most similar cases are neighbor. It counts the number of cases for each class and then assigns the new case to the same class to which most of its neighbors belong. This technique and its algorithm have been used for ages.

When one observes a decent person who becomes close to suspected criminals then by intuition, then he or she will cringe. This technique works well when we have identified a group; the chances are that any other persons among the group will likely be associated with them. Taking for instance a group of notorious armed bandits, when any person is seen near each member of the gang, our perception is that he or she is one. Another example is that in aplace notorious for its gang activities, any person wondering around is automatically assumed to be one of them.

The K-nearest neighbor counts the occurrences of cases "K" and assigns a new case to the highest number in the group. (Fig-22)

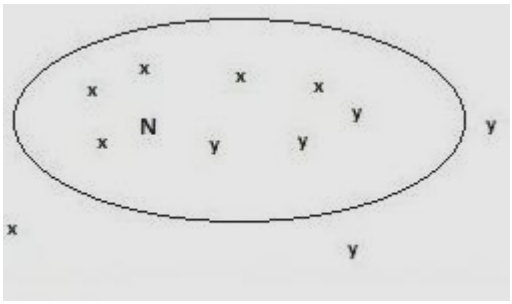


Figure 22: K-NN assigns the new case

The K- Nearest Neighbor assumes that when a crime is committed, in order to find the perpetrator look for a crime of similar pattern and try to find a pattern. Same is confined with this until no more similar cases could be matched with the cases as depicted in Fig-22.

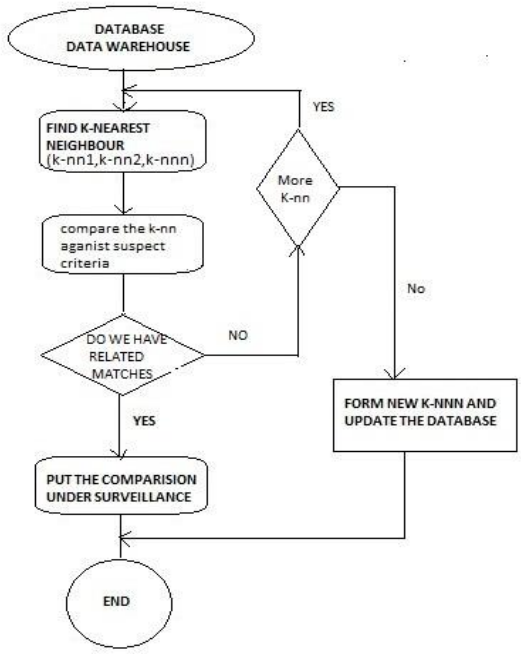


Figure 23: Application of K-NN Algorithm to identify crime cases

Mathematical representation of k nearest neighbor method

The trimming samples are described by n-dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor technique searches the pattern space for the k training samples that are closed to the unknown sample.

These k training samples are the k nearest neighbors" of the unknown sample."Closeness" is defined in terms of Euclidean distance, where the Euclidean distance between two points.,X=(x1,x2,…..xn) & Y=(y1,y2……yn) is

$$d\,(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The unknown sample is assigned the most common classes among its k nearest neighbors. When k=1 the unknown sample is assigned the class of the training sample that is closest to it in pattern space.

K- nearest neighbor algorithm

Input: Let U be the unknown tuple whose classes we want to assign,

Let T be the training set containing training tuples,

T1=(ti.1,t1.2,,…….t1.n),T2=(t2.1,t2.2…..t2.n)…….Tm=(tm.1,tm.2…..tm.n)

Let attribute ti1n be the class label of Ti

Let m be the number of training tuples

Let n be the number of attributes describing each tuple

Let k be the number of nearest neighbors we wish to find.

Output: Class label for U

Method: The method is outlined as follows

1. array a[m][2];// m rows containing data regarding the m training tuples.The first column is the Euclidean distance between U and that rows training tuple.The second column refers to that training tuples index. We need to save the index because when sorting the array (according to Euclidean distance), we need some way to determine to which training set the Euclidean distance refers.

2. For i=1 to m do

{

3 a[i][1]= Euclidian _distance(U1,Ti);

4 a[i][2]=I;}// save the index, because rows will be sorted later

5 sort the rows of a by their Euclidean distances saved in a[i][1] (in ascending order);

6. Array b[k][2];//The first column holds the distribute class labels of the k nearest

Neighbors, while the second holds their respective counts. In the worst case, each k nearest neighbor will have a different class's label, which is why we need to allocate room for k class labels.

7. For i=1 to k do{

8. If class label ta[i][2],n already exists in array b then

9. Find that class labels row in array b and increment its count;

10. Else add the class label in to the next available row of array b and increment its count;}

11. Sort array b in descending order (from class label with largest count down to that with smallest count);

12 Return (b[1]);// return most frequent class label of the k-nearest neighbor of U ass the class prediction.

Nearest neighbor technique are instance based or lazy learners in that they store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified. This controls with eager learning methods such as decision tree induction and back proportion, which construct a generalization model before receiving new samples to classify. Nearest neighbor classification can also be used for prediction ie to return a real valued prediction for a given unknown sample. In this case the classifier returns the average value of the real valued of labels associated with the k nearest neighbor of the unknown sample.

Proposed system

The web is used as infrastructure by terrorist groups for various aims. One examples is the forming of new local cells which may become active and perform acts of terror. The advanced Terrorist Detection system (ATDs), is aimed at combing down online access to abnormal content, which may include terrorist generated sites, by analyzing the content of information accessed by the web users. Further ATDs operate in two modes:

(i) The Training Mode

(ii) Detection Mode

(i) The Training Mode

It is first module of project where we can design terrorist transaction data base acknowledge their behavior from their web activities. As our database is prepared we will connect with our next module.

(ii) Detection mode

In detection mode terrorist behavior is given to detector module as reference data library. In this mode we calculate one threshold value, and content based detection. If we find such activity on web our system will make alarming reporting.

The proposed module for detection of terrorist is given below:

7.3.6.5 Multiparty source computation

The multiparty secure computation is a technique to identify the terrorists. In this technique the computation across multiple databases are done without revealing any information about data elements to other parties. Here it is assumed that the parties are semi honest i.e. they follow the protocol specification correctly /maximum when they attempts to extract extra information by the analysis of messages which are passed. We present here an example based on the article of Agarwal *et al.* (Agarwal R, 2003)

Agarwal R, Evfimievski, A, and Srikant, R (2003), "Information sharing across private databases". In proceedings of the 2003 ACM SIGMOD intl conf.on Management of data, San Diego,CA.

(Homeland insecurity: Data mining, Terrorism Detection, and Confidtiality by Stephen E. Fienberd Technical effort Number 148 Dec 2004)NISS.org

Let p and Q be two parties. They present a pair of encryption functions m(known to p only) and n (known to B only) such that for all y

M (n(y) =n (m(x))

The dataset of p consists of list P and the data set of Q consists of a list Q_.The message m(p)- is sent by P to q.q COMPUTES N(M(P_)) and then sends that to p the two information's n (m(p)) and n(q).p then applies m to computers N(M(P)u N(M(q)now A computers n(m(p))Un(m(Q).

Since p is knowing the order of items in P, p is also knowing the order of items in n (m(p) and can at once obtain PUQ.

The main disadvantages with this techniques are (1) it is asymmetric ie q must believe p to send PUQ back and

(i) It resumes semi honest behavior.

(ii) Functionality is paramount and privacy is only preserved to the content that the function outcome itself does not reveal information about the individual inputs.

(iii) In privacy preserving Data mining methods for terrorist detection is that they seek the protection of the latter while revealing individual records using the functionally of the format.

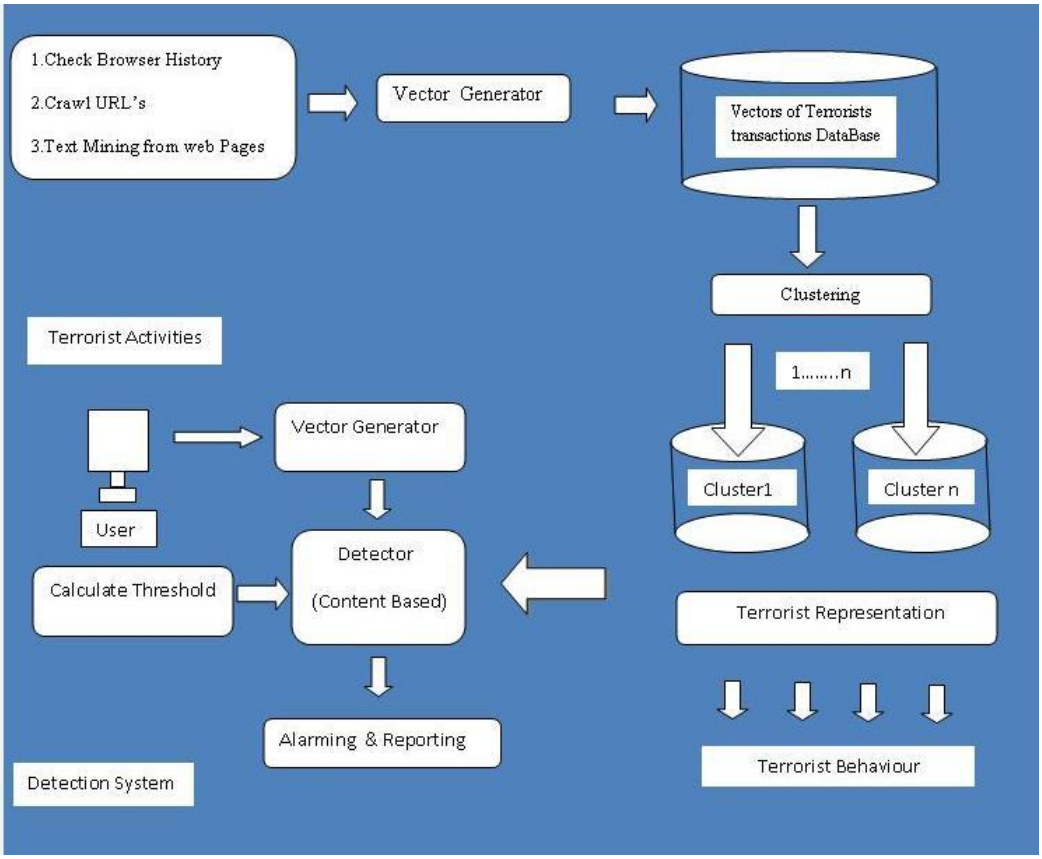


Figure 24: Detecting Terrorists Monitoring Module

7.3.7 Web pages related to Terrorists

Terrorists use web /net to communicate with each other and also to spread the terror.By using Data mining techniques we can recognize that usage of net. We
extract the log files over the net and check against some standard data bases to decide whether those URL's are suspicious or not (3).The following ways are:

(i) To check browser history

(ii) Crawl URL's and

(iii) Text mining from web pages.

Vector space model

We need to represent the content of terror related pages as against the content of currently accessed page in order to efficiently compute the similarity between them. This study uses the vector space model, a document is represented by an n dimensional vector d= (w_1, w2, w3…..w_n) where wi, represents the frequency based weight of term I in document d(5).The similarity of

two documents represented as vectors can be computed using of one of the known vector distance methods such as education distance or Cosime (2).

Transaction data base

We store vectors generated by vector generated by vector generator in database that is called transaction database. This database is further used for the clustering.

Clustering

Clustering is the process of partitioning data objects e.g. documents, records etc) into meaningful groups or clusters such that objects within a cluster have similar characteristics but are dissimilar to objects in other clusters. Cluster analysis can be observed as unsupervised classification of un labeled patterns (observations, data items or feature vectors) because no predefined category labels are associated with the objects in the training set.

Clustering results in a compact representation of large data sets (e.g. collections of visited web pages) by a small number of cluster controls.

Clustering applications include Data mining retrieval, image segmentation and pattern classification. Thus clustering of web documents viewed by internet users can revel collections of documents belonging to the same topic, clustering can also be used for anomaly detection

Representation of terrorists

Applying clustering algorithm from transaction data base n clusters is produced. These are used to represent terrorists behavior which is further used as input to the detector.

Detection systems

It is a new type of knowledge based detection methodology which uses the content of web pages browsed by terrorists and their supporters as an input to a detection process. In this study we only refer to the textual content of web pages excluding images, music, video clips and other complex data types. It has been assumed that terror related content usually viewed by terrorists and their supporters can be used as training data for a learning process to obtain a typical terrorist behavior. This typical behavior will be used to detect further terrorists and their supporters. A terrorist typical behavior is defined as an access to information relevant to terrorists and their supporters.

The components of detection syntax are mentioned below:

(i) User

(ii) Vector Generator

(iii) Detector(control based)

(iv) Alarming and reporting

Experimental results

Algorithm used WORD Extraction Algorithm

The word extraction algorithm is provided below:

Input: Web URL's i.e. log file

Output: Extracted meaningful word

Begin

Step1: Take the browser history

Step 2: Using vector extract the word from that document

Step3: match that word with standard data base.

Step 4: use clustering algorithm to do the cluster of similar words

Step 6: Report the problem using alarm or beep sound

Step 7: else the connection.

Exist

Monitoring mode

In this mode, the comparison of the content of information by the users and the typical terrorist behavior is made.

The textual content is represented in the form of vector called "access vector"[1].The clustering module access the collected vectors and performs unsupervised clustering resulting in n clusters representing the typical topics viewed by terrorist users. For each cluster the terrorist represent or module computes the centroid vector (denoted by vi) which represents a topic typically accessed by terrorists. As a result a set of centroid vector represent a set of terrorist's interests referred to as the typical terrorist behavior.

Text analysis is used to discover unknown, valid patterns and relation ships in large datasets. Even text analysis has a great potential to identifying unknown text documents, there is a limitation that human written languages is still complicated for machine to understand semantic meanings of it. over the years many studies have been made by using statistical methods to represent documents in to meaningful sequcences,such as TF-idF. This is most basic method for determine which words are significant in the given text data set. However performance it at based on tf-idf is not acceptable if thee amount of data is too small.morever this approach depends on the bag of words to calculate tf-idf value.

TF (t) = (n/N)

And idF(t)= $\log_e(\frac{T}{t})$

Where n= number of times term t appear in a document

N= total number of terms in document

T=total number of document

t= number of documents with term t in it.

The typical terrorist behavior is based on a set of web pages that were downloaded from terrorist related sites and is the main input of the detection algorithm. In order to make the detection logarithm move accurate the process of generating the typical terrorist behavior has to be repeated periodically due to changes in the content of terrorist related site.

Typical terrorist behavior depends on the number of clusters, when the number of clusters is higher typical terrorist behavior includes move topics of interest by terrorists where each topic is based on fewer pages.

It is hard to hypothesize what the optimal numbers of clusters are presented. The detector issues an alarm when the similarity between the access vector and the nearest centroid is higher than the predefined threshold denoted by tr:

$$Max\left\{\frac{\sum_{i=1}^{m}(tcv_{in}.tAv_i)}{\sqrt{\sum_{i=1}^{m}tcv_{12}^2.\sum_{i=1}^{m}tA_i^2}},\ldots,\frac{\sum_{i=1}^{m}(tcv_{in}.tAv_i)}{\sqrt{\sum_{i=1}^{m}tcv_{in}^2.\sum_{i=1}^{m}tA_i^2}}\right\}>t$$

Where cv is the ith centroid vector, Av the access vector,t cv the i^{th} term in the vector cv,tav-the i^{th} term in the vector av and m the number of unique t

Erms in each vector.

Advantages

- We can easily differentiate a normal user and user which uses net for harming the nation.
- Browsers history or web logs provide an exciting new way of collecting information on visitors.
- It provides better marketing intelligence

7.3.8 Results & Conclusions

It is evident from above result & discussing that Data mining techniques used in crime detection depends solely on the situation at hand. Most cases require the combinations of two or more techniques used along side for example, classifications & ink analysis technique can be used to complement each other.

An innovative knowledge based methodology for terrorist detection using web traffic content and using Data mining techniques has been presented in this chapter. The proposed methodology called ATDS learns the typical behavior (profile) of terrorists by applying a Data mining algorithm to the textual content of terror related websites. The resulting profile consisting

of logs is used by the system to perform real time detection of users suspected of being engaged in terrorist activities. The Receiver Operator Characteristic (ROC) analysis shows that this methodology can outperform a command based intrusion detection system.

The results suggests that the methodology for terrorist activity detection in the web can be very w useful to detect terrorists and their supporters using a legitimate ways of internet access to view terror, related content at a series of evasive web sites. Present system just recognize the words of terrorists languages.by developing such system,relation ship between human and computer becomes much closer and secure. The proposed approach is quite efficient to detect terror related activities.

Data mining is not all to counter terrorism since there are a large no of drawbacks which included the problems of skilled man power, in inadequate investment in it & telecommunication infrastructures in adequate Data mining policies & above all legal issues that characterize unwanted tracking of in on centric citizens.

Combating terrorism, terror attacks and other terror related activities require the proper attention of the government, law enforcement agencies should be better equipped in this communication edge on how to use computer & computer analysis to track the nefarious activities of the hoodlums corporate bodies particularly backs should also play vital role in the fight against terrorism related activities.

7.3.9 Privacy Preserving Data Mining

Privacy: Firstly we briefly mention the challenges due to Data mining, there has been much debate recently among the counter terrorism experts and civil liberties unions and human rights lawyers about the privacy of individuals. That is gathering information about people, mining information about people, conduction surveillances activities and xa mining say & email messages and phone counter stations are all threats to privacy and civil liberties. The question arises how can we maintain the privacy but at the same time ensure the safety of nations from terrorism. What should we sacrifice and to what extent?

7.3.10 Techniques for privacy preservation

A number of techniques have been suggested for privacy preservation.

7.3.10.1 Authority control & Cryptographic techniques

Such techniques effectively hide data from unauthorized access but do not prohibit/show in appropriate use by authorized users (Pinkas, 2002).

7.3.10.2 The anonymisation of the data

In which any identifying attributes are removed from the source data sets. Various on this can be after applied to the rule set to suppress rules containing identifying attributes.

7.3.10.3 Query Restriction

Which attempts to detect when statistical comprise might be possible through the combination of queries (Miller, 1991, Miller & Seberry, 1989)

7.3.10.4 Dynamic sampling

Dynamic sampling and reducing the size of the available data sets. This can be done by selecting different sets of a source tuples for each query.

7.3.10.5 Noise addition & Data perturbation

If individual's entries in such a manner as to retain the accuracy of statistical queries. This can be done in the following two ways:

(a) Noise addition in which sets of values are changed such that common statistical & mining operations yield the same result. (Agarwal & Srikant 2000)

(b) A randomization techniques with a similar effect is suggested by (Evfimievski 2000).

7.3.10.6 Data swapping

Where attribute value is interchanged in a way which maintains the result of statistical queries (Evfimievski 2000)

Multiparty computation: Clifton it al discuss four methods in which multiple sites can generate rules without compromising each sides data. (Clifton it al. 2002)

7.3.11 Future direction

Changing detection methodology more accurate results may be obtained by monitoring sequences of page views rather than raging an alarm after every suspect page. Developing an anomaly detection system for detecting abnormal content, which may be an indicator of terrorists or other criminal activities, is another important research direction which has been started by Cast *et al.* (2001).

Moreover applying classification methods to the terrorist detection problem is another interesting future direction.

Changing Computational Complexity

A system based on the new methodology has to process every HTML page. Which is being accessed by any subscriber of an ISP where it is deployed? There is a need to work on reducing the computational complexity of the proposed methodology, one way to obtain this goal is to minimize the size of each access vector (Dimensionality reduction) without significantly reducing the system detection performance.

Optimal settings

Further analysis is needed to obtain the system settings such as the number of clusters (k) and the detection threshold.

REFERENCES

Abbasi, M. M.& Kashiyarndi S., Clinical Decision support system, discussion on different methodologies used in healthcare (2006).

Abdullah A. Aljumah., Ahamad Mohammad Gulam. Siddiqui Mohammad Khubeb, Application of Data mining: Diabets health care in young and old patients, Journal of King Saud University Computer & Information Science 25, pp.127-136, (2013).

Agarwal R., Ghosh A., Imilinski, T. Iyer B. & Swami A. An interval classifier for database mining applications in proceedings of the 18th conference on very large databases, Morgan

Agarwal, R, Efimievski, A. and Srikant, R. Information sharving acress private Databases in proceedings of the 2003 ACM SIGMOD intel.conf on management of data, San Diego, A (2003).

Breault, Joseph L., Goodall Colin R., Fose Peter J., Data mining a diabetic data warehouse Artifical Intelligence in Medicine 26, pp-37- 54. (2002).

Chen M. S., Han J. and Yu P.S. Data mining: An overview from database perpective IEEE, transactions on knowledge and Data Engg, 8(6):866-883 (1999).

Chiang Yu-Min, Yu Chieh Lo and Shang-Yilin The application of Data mining technique and Multiple classifiers to marketing Decision, International Journal of Electronic Business Management, Vol-3, No-4, pp. 301-310 (2005).

Chirsty T. Analtytical tools help health firms, Fig-ht, fraud, insurance & Technology, 22(3), pp.22-26, (1997).

Dey Monali, Ratanargy Siddarth Swarup. Study and Analysis of Data mining algorithms for Healthcare decision support system, IJCSIT, pp. 470-477, Vol-5 No-1, (2014).

Faryal G., Wasi B.H. and Usman Q. Terrorist group prediction using data classification, Presented at the international conferences of Artificial Intelligence and pattern Recognition Malaysia, (2014).

Fayyad Usman, Gregory Piatetsky Shapiro and Padhraic Smyth. From Data mining to knowledge discovery in databases, American association for artificial Intelligence/The MIT press. (1996).

Gillespie G. There's gold in them that' databases, Health Data Management, 8(11), 40-52. (2000).

Han Jiawei Data Mining Concept and Techniques

Han, J., Kamber, M Data mining: Concepts and techniques, Morgan Kaufram (2001).

Hian Chye kob and Tan Gerald, Data mining applications in healthcare Journal of healthcare Information management, Vol-19, No- 2, pp.64-71 (2011).

Johnson, D.E.I Web based data analysis tools help providers, MCOs contain costs.Healthcare Strategic Management, 19(4), pp.16-19. (2001).

Kavifinanpubs (LosAltos CA), Vancouver (1992).

Khan Muhammad Naeem, Ahmed, Qureshi Sheraz Ahmad & Riyaz Naveed, Gender Classification with decision trees, International Journal of Signal processing image processing and pattern recognition, Vol-6, No-1 Feb (2013).

Kincade K., Data mining: Digging for healthcare gold, Insurance & Technology, 23(2), IM-IM7.(1998).

Kolar, H. R. Caring for healthcare, health management technology, (22-14),46-47. (2001).

Kreuze D. Debugging hospitals technology review, 104(2), 32. (2001).

Lilian Sing'oei & Jiayang wang, Data mining framework for direct marketing A case study of Bank marketing by International Journal of Computer Science issues, Vol-10, Issue-2, No-2, March, (2013).

Memon nasulla and Abdul qureshi investigative data mining and its application in counter terrorism proceeding of the 5 international conference on applied information and communication,pp.397-303(2005).

Millay A. healthcare Data mining, Health management technology21 (8)pp. 44-47, (2000).Ingram M.,Internet privacy threatened following terrorist attacks on VS,URL : http://www.wsws.org/articles/2001/Sep 2001/isps24, shtml. (2001).

Miller Ramdolph A.,medical diagonessstic decision support systems past present and future,a threshold bibliography and brief commentary, journal of the American medical Information association 1, 1 pp.8- 27.(1994).

Mulley Pradyna. & Joshi Anniruddha., Application of Data mining technologies for customer Segmentation in real time business intelligence International Journal of Innovative Research in Advanced Enginnering, Issue-4, Vol-2,pp.23-49, 21-63 (2015).

Petre Ruxandra Data mining Solution for the business environment by, Database system Journal Vol-4,No-4,pp.21-29 (2013).

Piazza P. Health alerts to Fig-ht biotterror, security management 46(5) 40. (2002).

Prabhu, S., N. Venkatesan Data Mining and Warehousing, New Age International(P) Limited

Pujari, A. K. Data Mining BPB Publications

Radhakrishanan B., Shineroj G,Anver Muhammad K.M,Application of Data mining in marketing IJCSN International Journal of Computer Science &Network, Vol- 2,Issue 5, pp.41-46. Oct (2013).

Reza fadaei-tehrari, Themas M. Green Crime and society International Journal of Social Economics Vol 29,MP 10,pp.781-795,(2002).

Schurenberg, B.K.An information excavation health data management 11(6),80-82.(2003).

Sharma Gajendra, Data Mining, Data Warehousing and OLAP by S. K. Kataria & Sons

Sharma Hemant Kumar, Dr. Sarmistha, A study of Data mining activities for market research, ZENITH, International Journal of Multidisciplinary Research Vol-1, Issue 8, Dec. ISSN 2231-5780, pp.205-212 (2011).

Silver M, SakataT., herman SU.H.C., dolins, S. B &o'Shev,M- 5 Case study, how to apply Data mining techniques in a healthcare data ware house, journal of healthcare information management,15(2),pp.155-164(2001).

Swami A.C. & V. Jain System analysis & design, pp. 324-333. (2009).

Takagi Noboru, An application of binary decision trees to pattern Recognetion Journal Advanced computational intelligence and intelligent Informatics, Vol-10,No-5 (2006).

Thurassingham Bhavani, Data mining for counter terrorism, pp.191-215, (2008).

Turcinkova, J., Stejskal, L., Stavkova, J. Chovani a rozhodovani spotrebitele consumer behavior and decision making Brno: MSD.104 pp.ISBN 978-80-7392-013-5.(2007).

Vighne S., Trimbark Priyanka, Musmade Anjali, Merukar Ashwini, Pandit Sandip, An approach to detect terror related activities on net IJARIIE, ISSN (o) 2395-4396, Vol-2, Issue-1, pp.401-406 (2016).

Zikmund W. G., M. C. Lolead, R. and Gilbert, F. W, Coustmer relationship management integrating marketing strategy and information technology,Jonvillc & sons INC,Nework (2003).

Data Mining Techniques & Applications

(ISBN: 978-93-88901-29-1)

About Author

Dr. Satish Chandra Pandey

Dr. Satish Chandra Pandey completed his Masters in Computer Science from Barkatullah University Campus, Bhopal & Ph.D. in Computer Science from Sunrise University, Alwar (Rajasthan) in 2019.His Main Research Work focuses on, Network Security, Hash Functions, Data Mining, and Computational Intelligence based education. He has supervised one Research Scholar and 2 are going on. He has published more than 17 research papers in reputed National & International Journals & Conferences including IEEE & CSI. & More than 75 National & International Webinars & FDP's have been also attended by him. He is a recipient of Best Young Faculty Award, 2022-23, by the Novel Research Academy for his academic Excellence Journey. He is a Life member of Computer society of India (CSI) (since 2008) & India Science Congress Association ISCA (since 2010). He is a member of International Association of Engineers (IAENG) & Asia Society of Researchers (ASR) also from 2020. Presently He is working as an Associate Professor in Computer Science & Engineering Deptt in Jayoti Vidyapeeth Women's University, Jaipur (Rajasthan) & Total Teaching Experience is more than 12 Years.

Bhumi Publishing

ISBN 978-93-88901-29-1
9 789388 901291