CONCEPTS, TECHNIQUS, APPLICATIONS AND ISSUES OF DATA MINING

By
S. C. Pandey, Lecturer, S.V College of Sc. & Hr. Edu, MIA, Alwar (Raj),
D.Dubey, Lecturer, Institute of Technology & Management, Gorakhpur (UP)
P.K. Singh, Director(P&T), S.V College of Sc.& Hr. Edu, MIA, Alwar(Raj),

Email: pandey.satishchandra@gmail.com

ABSTRACT: Data mining is the techniques of extracting useful information from large and unorganized data banks. It is the process of performing automated exploration and producing predictive information from large databanks. It is the process of finding previously unknown patterns and trends in databases and using information to build predictive models. It enables us to understand the current market trends and enables us to take proactive measures to gain maximum benefit from the same. In this paper concept, various techniques of data mining, applications of data mining in marketing and various issues of data mining have been discussed and analyzed.

Introduction:

Data mining can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown pattern [1].It aims to identify valid, novel, potentially useful and understandable correlations and patterns in data by combing through copious data sets to sniff out patterns that are too subtle are complex for humans to detect [2]

Several factors have motivated the use of data mining applications in healthcare. The existence of medical insurance fraud and abuse for example hassled many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders [3].

Another factor is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision making by discovering patterns and trends in large amounts of complex data [4].

Data mining applications can be developed to evaluate the effective ness of medical treatments .By comparing and contrasting causes, symptoms and courses of tretements, data mining can deliver an analysis of which courses of action prove effective [1].For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost effective [5].

Johnson [6] has suggested that at a higher level, data mining can facilitate comparisons across healthcare groups of things such as practice pattrns, rsource utilization, and length of stay and costs of different hospitals.

Data mining can be used to detect the terror related activities on the web[7].It is assumed that terror related content usually viewed by terrorists and their supporters can be used as training data for a leaning process to obtain a ' Typical-Terrorist

Behviour.'This typical behavior will be used to detect further terrorists and their supporters. A terrorist typist behavior is defined as an access to information relevant to terrorists and heir supporters. In the above suggested methodology each user under surveillance is identified as a user computer having a unique IP address rather than by his or her name. In the case of a real time alarm, the detected IP can be used to locate the compute and hopefully the suspected terrorist who may still be load onto the same computer.

In marketing data mining is becoming increasingly popular of no increasingly essential. More details on data mining techniques can be found in berry and Linoff [8].Most of the industries need to advertise and promote their products and services. Insurance companies, banks and retail stores are typical examples. Her are generally to way to advertisement and promotion (i) mass marketing and (ii) direct marketing. Mass marketing uses mass media such as raio, transistor, elevision and newspapers broad casts messages to the public without discrimination. This was an effective way of promotions hen the products were in large demand by the public. But no a days where products are tremendous in numbers and the market is highly competitive mass marketing is not so effective.

The response rate, the percent of people who actually purchase the products after seeing the promotion is often low.

The second way of promotion is direct markting.In place of promoting the customers indiscriminatively.In direct marketing study customs characteristics and needs and select some customs as the target for promotion. The hope is that the response rate for the selected customers can be much improved. At present a large amount of information on customers is kept in data bases hence data mining can be very effective for direct marketing.

Data mining has been widely used in direct marketing to target customers by a number of researchers [9, 10,11].Data mining has also been defined by Fayyad, Piatetsky-Shapiro,Smyth and Utturusammy[12] as a nontrivial process of discovering novel, implicit useful and comprehensive knowledge from a large amount of data. In direct marketing this knowledge is a description of likely buyers and is useful in obtaining higher profit than mass marketing.

Process of Data Mining:

(i) The various steps [13] in the data mining process to extract useful information are:

(ii) Problem definition: This phase is to understand the problem and the domain environment in which the problem occurs. We need to clearly define the problem before proceed further. The Problem definition specifies the limits within which the problem needs to be solved. It also specifies the cost limitations in solving the problem.

(iii) Creation of a database for data mining: This phase is to create a database where the data to be mined are stored for knowledge acquisition. The creation of data mining database consumes about 50% to 90% of the overall data mining process. Data warehouse is also a kind of data storage where large amount of data is stored for data mining.

(iv) Searching of the 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.

(v) Creation of a 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.

(vi) Building a 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.

(vii) 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.

(viii) Deployment of the 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. The following fig. shows the various phases in the data mining process.

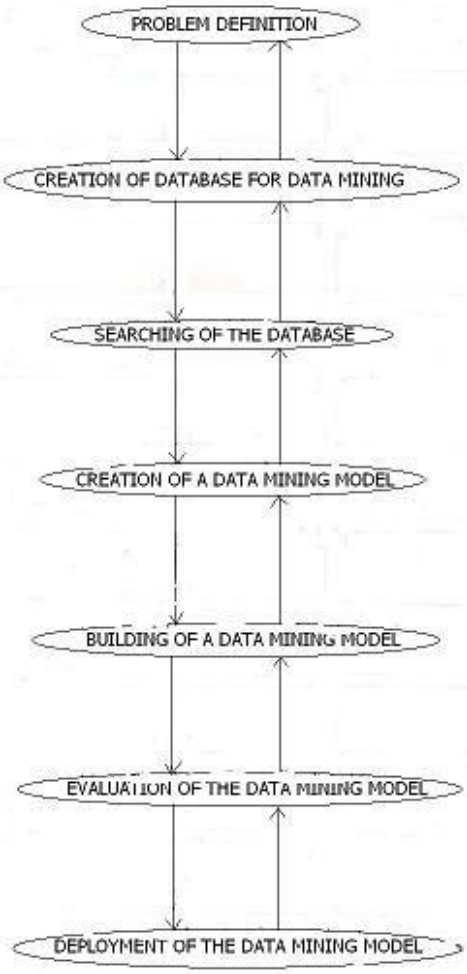

FIG-1 MAIN STEPS OF DATA MINING PROCESS

… ……

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[13] have been proposed by various individuals and organizations that provide systematic phases for data mining. The four most popular process models of data mining are:

(a)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 are related to the problem. After that we analyze the accumulated data from different angles using various data mining techniques. We then extract meaningful information from the analyzed data and implement the result in solving the problems 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 shows the life cycle of the 5A's process model.

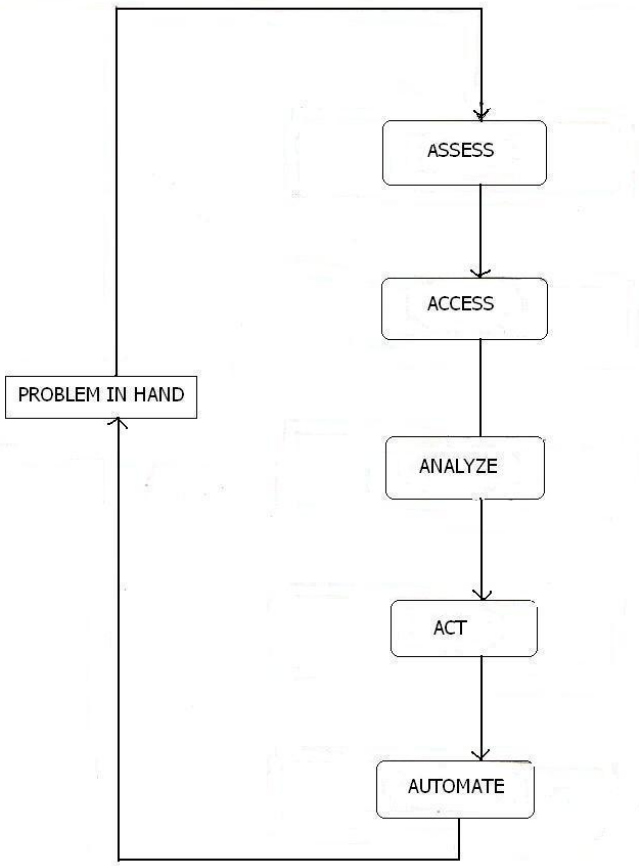

FIG-2 LIFE CYCLE OF 5A'S PROCESS MODEL

(b)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:

(i) Understanding the business: This phase is to understand the objectives and requirements of the business problems and generating a data mining definition for the business problem.

(ii) 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 hypo these for solving the problem.

(iii) Preparing the data: This phase is to create final datasets 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.

(iv) Modeling: This phase is to select and apply different modeling techniques of data mining. We input the data sets collected from the previous phase to these modeling techniques and analyze the generated output.

(v) Evaluation: This phase is to evaluate a model or a set of models that we generate in the previous phase for better analysis of the refined data.

(vi) Deployment: This phase is to organize and implement the knowledge gained from the evaluation phase in such a way that it is easy for the end users to comprehend.

(vii) The following fig. shows the life cycle of the CRISP-DM process model.

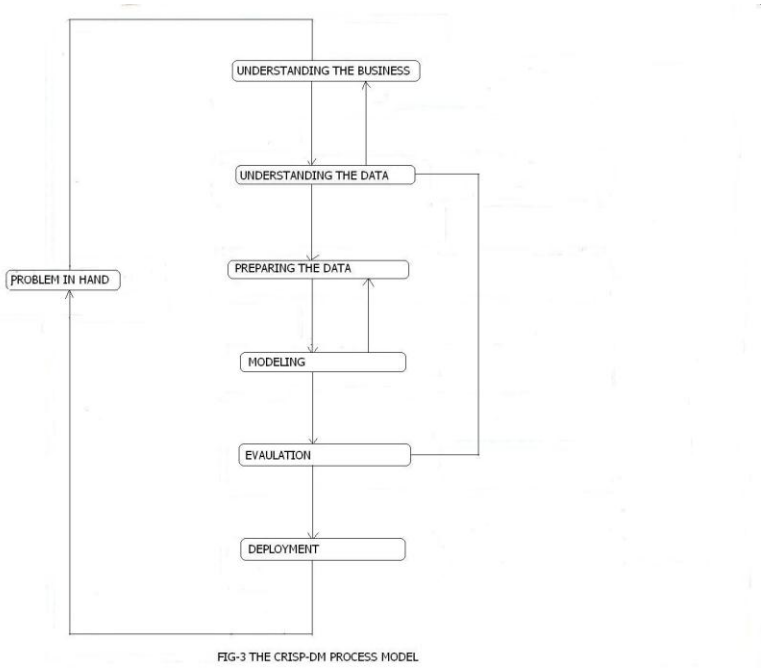

FIG-3 THE CRISP-DM PROCESS MODEL

(c) The SEMMA process model: In this process model SEMMA stands for Sample, Explore, Modify, Model and Assess. The life cycle of the SEMMA process model consists of five phases.

(i) 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.

(ii) 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.

(iii) 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.

(iv) 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 required result for the problem.

Some of the modeling techniques that we can use a model are neural network and statistical models.

(v) 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 estimates its performance. The following fig shows the lif cycle of the SEMMA process model.

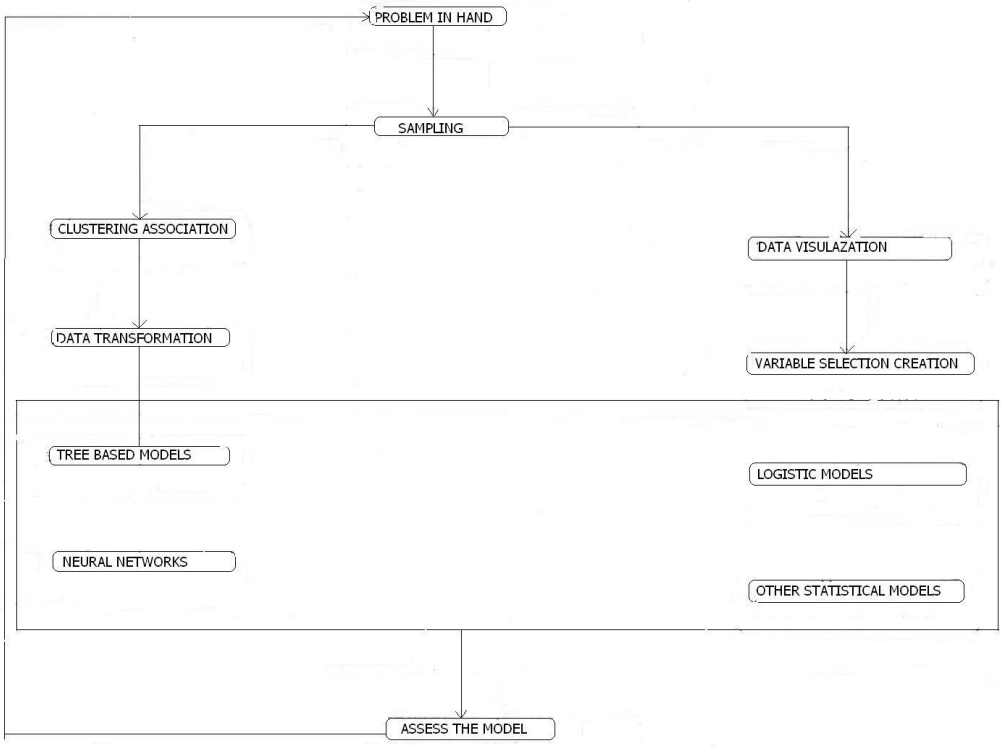

FIG-4 THE SEMMA PROCESS MODEL

(d) The six sigma process model: The six sigma 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 easy implementation and it is likely to be implemented world wide. 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, Analyses, Improve and Control.

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

(i) Define: This phase is to define the goals of a project along with its limitations.

(ii) Measure: This phase is to collect information about the current process in which the work is done and to try to identify the basis of the problem.

(iii) Analyze: This phase is to identify the root cause of the problem in hand and ensure those root causes by using various data analysis tools.

(iv) Improve: This phase is to implement all those solutions that tries and solves the root causes of the problem in hand.

(v) Control: This phase is to monitor the outcome of all its previous phases and suggest improvement measure in each of its earlier phases. The following fig shows the life cycle of the six-sigma process model.

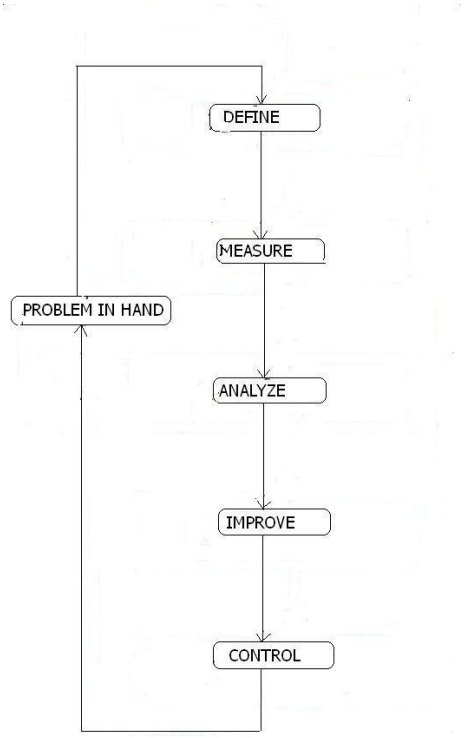

FIG-5 TH SIX SIGMA PROCESS MODEL

Data Mining Techniques:

The most commonly used techniques [14] in data mining are:

(a) **Artificial Neural networks:** Non linear predictive models that are learnt by training & resemble and variable biological neural networks in structure.

(b) **Decision trees**: The decision tree methods include classifications and regression tress (CART) and chi-square automatic interaction detection (CHAID). These resemble tree shaped structures.

(c) **Genetic algorithms**: Optimizations techniques that use process such as genetic combination, mutation and natural selection in a design based on the concepts of evolution.

(d) **Nearest neighbor method**: A technique that classifies each record in a data set based on a combination of the classes of the k record(s) most

similar to it in a historical dataset. It is sometimes called the k-nearest neighbor technique.

(e) **Rule Induction**: The extraction of useful if then rules from data based on statistical significance.

(f) **Data Visualization**: The visual interpretation of complex relationships in multidimensional data ,graphics tools are used to illustrate data relationships.

How Data Mining is applied:

The technique that is used in data mining is called modeling. Modeling is simply the act of building a model in one condition where one knows the answers and then applying it to another condition that one does not know.

The following two examples may help in understanding the use of data mining for building a model for new customer.

Let us suppose that the marketing director has a lot of information about his prospective customers i.e. their age, sex, credit history etc.

Now his problem is that he don't know the long distance calling usage of these prospects (because they are most likely now customers of his competition).He would like to concentrate on those prospects who have large amounts of long distance usage. Then he can obtained this by building the model [8] as shown in the table-1

Table-1 data mining for Prospecting

	Customers	Prospects
General Information(e.g. Demographic data)	Known	Known
Propriety any information(e.g. Customers transaction)	Known	Target

The aim in prospecting is to make some calculated guesses about the information in the lower right hand quadrant

Which is based on the model that he builds while going from customers general information to customers proprietary information.. With this model in hand new customers can be selected as target.

Another common example for building the models is shown in the table-2.Test marketing is an excellent source of data for this kind of modeling. By mining the results of a test market which represents a broad but relatively small sample of prospects one can provide a base for identifying good prospects in the overall market.

Table2 Data mining for predictions

	Yesterday	Today	Tomorrow
Static information and current plans(e.g. demographic data, marketing plans)	Known	Known	known
Dynamic information(e.g. customer transactions)	Known	Known	target

Data mining can be used for direct marketing to get higher profit as compared to mass marketing. For that whole database contains 500000 customers. In direct marketing only 20% of the customers identified as likely buyers by data mining (which costs Rs. 100000) are chosen to receive the promotion package in the mail. The mailing cost is thus reduced dramatically. At the same time however the response rate can be improved from 1% in mass mailing to 3 %(real improvements for a 20% rollout).
We see from table-3 that net profit from the promotion becomes positive in direct mailing compared to a loss in mass mailing.

Table-3 A comparison between directional campaign and mass mail campaign.

Details	**Mass mailing**	**Direct Mailing**
Number of customers mailed	500000	(10%) 50000
Cost of printing, mailing(Rs 60.00 each)	30000000	3000000
Cost of data mining	Nil	1000000
Total promotion cost	30000000	4000000
Response rate	1.0%	3.0%
Number of sales	5000	1500
Profit from sale(Rs. 4000 each)	20000000	6000000
Net profit from promotion	-10000000	2000000

Results & Discussions: In the example of the direct mail campaign using data mining as illustrated in tables3.We can clearly see that if we mail only to some top small percent of customers even then the net profit from the promotion becomes positive in direct mailing as compared to a loss in mass mailing. Hence using data mining the return of investment or net profit can be improved.

Thus we demonstrated that data mining is an effective tool for direct mining which can bring more profit to banks, insurance companies and the retail industry than the traditional means of mass marketing.

Data mining issues:

Although Data mining has been developed a conventional ,mature,trusted and power ful technique even then these are certain problems[15] related to data mining which are discussed below in detail.One should note it that these problems are not exclusive and are not ordered in any way.

(i) **Issues related to security and social matters**: Security is an important problem with any type of data collection which is shaved and/or is intended to be used for strategies decision making [9].Moreover when data is collected for customers profiling, user behavior understanding, correlating personal data with other information etc .large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controvertisial given the confidential nature of some of this data and the potential illegal areas to the information. Moreover data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies.

(ii) **Issues related to user interface**: The knowledge invented by data mining tools is useful as long as it is interesting and understandable by the end user. Good data visualization simplifies interpretation of data mining results as well as helps end users to better understand their needs. The major problems related to user interfaces and visualization are 'Screen real estate', information rendering and interaction.Intacterivity with stored data and data mining results is essential because it provides means for the end user to focus and refine mining tasks and to visualize the discovered knowledge from different angles and at different conceptual levels.

(iii) **Issues related to performance**: Artificial intelligence and statistical methods for the data analysis and interpretation are generally not designed for mining large data sets. Data sets size in terabytes is common now a day. As a result, processing large data sets raises the problems of scalability and efficiency of data mining methods. It is not possible to fractionally use algorithms with exponential and even medium order polynomial complexity for data mining. Linear algorithms are generally used for mining large data.

(iv) **Issues related to data mining methods**: It is often desirable to have different data mining methods available because different approaches mine data differently depending upon the data in hand and the mining requirements [9].The algorithms that we use in data mining assumes that the stored data is always noise free and in most of the cases it is a forceful assumption. Most data sets contain exceptions, invalid or incomplete information which complicates the data analysis method. Presence of noisy data reduces the accuracy of mining results. Due to which data preprocessing i.e. the cleaning of data and its transformation becomes essential. Data cleaning is a time consuming process but it is one of the most important phase in knowledge dis

(v) **Issues related to data source**: Data mining systems rely on databases to supply the raw data for input and this raises serious problems because databases are dynamic, incomplete, noisy and large[10] The current trend[9] is to collect as much of data as possible and mine them later as and when required. The concern is about the quality and type of the large data being collected;

very clear understanding is required to collect right data of proper amount and to distinguish between useful and useless data. Now a day's databanks are of different types and stores data with complex and diverse data types. It is very difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different data types and data sources require specialized mining algorithms and techniques.

Data Mining Applications:

Data mining applications in news & Entertainment data analysis & management: News and Entertainment industry generate large amount of data in the form of text, graphics audio and video formats which are read and viewed by people who are demographically different from each other. Data mining enables the news and entertainment organizations to study and analyze such diverse collection of data and retrieve meaningful and useful statistics that can be implemented in future for better readership or viewer ship of these programs.

(i) Data mining applications in **telecommunications analysis and management**: Data mining enables us telecommunication analysis to consolidate telecommunication set up by providing them with reduced cost of doing the business, improving profit and enhancing the quality of service to the consumers.

(ii) Data mining applications in **scientific and Engineering data analysis and management**: Data mining models scientists and engineers to use various statistical techniques to analyze scientific and engineering data.

(iii) Data mining applications in **stocks and investments analysis and management**: Data mining enables us to study fist the specific patterns of growth or downslides of various companies and then intelligently invest in a company which shows the most stable growth for a specific period.

(iv) Data mining applications in **crime analysis and management**: Data mining enables security agencies and police organizations to analyze the crime rate of a city or a locality by studying the past and the current trend that led to the crime and prevents the reoccurrence of such incidences and enables concerned authorities to take preventive measures.

(v) **Data mining applications in computer security analysis and management:** Data mining enables network administrators and computer security experts to combine its analytical techniques with our business knowledge to identify probable instances of fraud and abuse that compromises the security of a computer or network.

(vi) **Data mining applications in insurance sector:** Data mining can help the insurance companies to predict which customers with buy new policies and can also identify the behavior patterns of risky customers and fraudulent behavior.

(vii) **Data mining applications in banking analysis and management:** Bank authorities can be able to study and analyze the credit patterns of their

consumer and prevent any kind of bad credits or fraud detection in any kind of banking transactions using data mining. By data mining the bank authorities can find hidden correlations between different financial indicators and can identify stock trading rules from historical market data. By data mining bank authorities can identify to change credit card affiliation.

(viii) **Data mining applications in marketing:** Data mining provides the marketing and sales executive with various decision support systems that helps us in consume acquisition, consumer segmentation consumer retention and cross selling. In this way it enables us to better interact with consumers, improve the level of consumer services that we provide and establish a song lasting consumer relationship. One can demonstrate that data mining is an effective tool for direct marketing which can give more profit to the retail industry than the traditional means of mass marketing.

(ix) **Data mining applications in healthcare analysis and management:** Healthcare organizations and pharmaceutical organizations provides huge amount of data in their clerical and diagnostic activities. Data mining enables such organizations to use the machine learning techniques to analysis healthcare and pharmaceutical data and retrieve information that might be useful for developing new drugs. When medical institutions use data mining for their existing data they can discover new, useful and potentially life saving knowledge that otherwise remained inert in their database.

REFERENCES:

[1] A.Milley (2000), Healthcare and data mining, Health Management Technology, 21(8), 44-47.

[2] D.kreuze(2001).Debugging hospitals.Technology Review,104(2),32.

[3] T.christy(1997).Analytical tools help health firms fight fraud.Insurance & Technology,22(3),22-26.

[4] SBiafore, (1999).Predictive solutions bring more poer to decision makes, health management technology, 20(10), 12-14

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

[6] D.E.I Johnson (2001).Web based data analysis tools help providers,MCOs contain costs.Health care Strategic Management,19(4),16-19

[7] J.Kelley (2002) Teror groups behind web encryption , USA Today, URL: http://www.apfn.org/apfn/wtc why.htm

[8]M.J.A.Berry & G.S. Linoff(1997).Data Mining Techniques:For Marketing,Sales and Customer Support.New York: John Wiley & Sons Inc.

[9] R. Agrawal, A.Ghosh, T. Jmielinski, B.Lyer, & A.Swami(1992). An interval classifier for database mining applications.In proceedings of the 18th conference on very large databases, Morgan Kaufinan pubs (Los Altos CA),Vancouver

[10] T.Terano, Y.Ishino (1996).Interactive knowledge discovery from marketing questionnaire using simuated breeding and inductive learning methods. In proceedings of the second international conference on knowledge discovery and data mining pp .279-282.

[11] V.Ciesielski,G.Palstra(1996).Using ahybrid neural/expert system for database mining in market survey data.In Proceedings of the second Interntional Conference on Knowledge Discovery and Datamining,pp.38-43.

[12] U.M. Fayyad,G. Piatetsky-Shapiro,P.Smyth,& R. Uthurusamy.(Eds).(1996). Advance in knowledge Discovery and Data Mining.MII press,Mento Park

[13] Data Mining, (BPB publications,B-14 ,Connaught place, New Delhi-1) p-15-16.

[14] System Analysis & Design by A.C.Swami & V.Jain(College Book House p. ltd. Chaura Rasta,jaipur ,p-328-329)

[15] G. Sharma, Data mining, data warehousing And olap (S.K.Katariya & Sons, Ansari road, dariyaganj, New Delhi) pp. 14-15