International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
Recent Trends in Mobile & Cloud Computing
(NCRMC- 08th & 09th October 2015)

RESEARCH ARTICLE OPEN ACCESS

Data mining in healthcare

Miss. Swati Raju Shinde
Computer Science & Engineering College of Engineering & Technology, Akola
swatishinde335@gmail.com

ABSTRACT
Data mining or knowledge discovery in database, as it is also known, is the known travel extraction of implicit, previously unknown and potentially useful information from the data. Data mining has been used intensively and extensively by many organizations. In healthcare data mining is becoming increasingly popular, if not increasingly essential.
KEYWORDS: Knowledge Discovery in Databases, Data mining, Data Mining Framework.

I. INTRODUCTION

The purpose of data mining is to extract useful information from large databases or data warehouses. Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns. Data mining is not new—it has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Several factors have motivated the use of data mining applications in healthcare.

The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders. Fraud detection using data mining applications is prevalent in the commercial world, for example, in the detection of fraudulent credit card transactions. Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection.

A.Knowledge Discovery in Databases (KDD)

The process of discovering useful knowledge from a huge data is called as Knowledge Discovery in Database (KDD) and which is often referred to as Data mining. While data mining and knowledge discovery in databases are normally treated as synonyms, but, in fact data mining is a part of knowledge discovery process. The KDD process comprises of few steps as shown in Fig. 1

Figure 1.1 Data Mining is the core of Knowledge Discovery process

Data collected from multiple sources often heterogeneous is integrated into a single data storage called as target data. Data relevant to the analysis is decided on and retrieved from the data collection. Then, it is pre-processed and transformed into an appropriate standard format. Data mining is a crucial step in which intelligent algorithm/techniques are applied to extract meaningful pattern or rules. Finally, those patterns and rules are interpreted to new or useful knowledge or information.

II. LITERATURE REVIEW

A literature review is a text written by critical points of current knowledge including substantive find theoretical and methodological contributions to a particular topic. Literature reviews are secondary sources and do not report any new or original experimental work.

Hian Chye Koh and Gerald Tan mainly discusses data mining and its applications with major areas like Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management.

Jayanthi Ranjan presents how data mining discovers and extracts useful patterns of this large data to find observable patterns. This paper demonstrates the ability of Data mining in improving

the quality of the decision making process in pharma industry. Issues in the pharma industry are adverse reactions to the drugs.

M. Durairaj, K. Meena illustrates a hybrid prediction system consists of Rough Set Theory (RST) and Artificial Neural Network (ANN) for dispensation medical data. The process of developing a new data mining technique and software to assist competent solutions for medical data analysis has been explained. Propose a hybrid tool that incorporates RST and ANN to make proficient data analysis and indicative predictions. The experiments onspermatological data set for predicting excellence of animal semen is carrie out. The projected hybrid prediction system is applied for pre-processing of medical database and to train the ANN for production prediction. The prediction accuracy is observed by comparing observed and predicted cleavage rate.

K. Srinivas, B. Kavitha Rani and Dr. A. Goverdhan discusses mainly examine the potential use of classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Using an age, sex, blood pressure and blood sugar medical profiles it can predict the likelihood of patients getting a heart disease.

Shweta Kharya discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis decision tree is found to be the best predictor with 93.62% accuracy on benchmark dataset and also on SEER data set.

Elias Lemuye discussed the AIDS is the disease caused by HIV, which weakens the body's immune system until it can no longer fight off the simple infections that most healthy people's immune system can resist. Apriori algorithm is used to discover association rules. WEKA 3.6 is used as the data mining tool to implement the Algorithms. The J48 classifier performs classification with 81.8% accuracy in predicting the HIV status.

Arvind Sharma and P.C. Gupta discussed data mining can contribute with important benefits to the blood bank sector. J48 algorithm and WEKA tool have been used for the complete research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9%.

III. WHAT IS DATA MINING?

Data mining can be considered a relatively recently developed methodology and technology, coming into prominence only in 1994. 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 or complex for humans to detect.

Data mining or knowledge discovery in database, as it is also known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes, and detecting anomalies.

A.Development of Data Mining:

The current evaluation of data mining functions and products is the results of influence from many disciplines, including databases, information retrieval, statistics, algorithms, and machine learning.

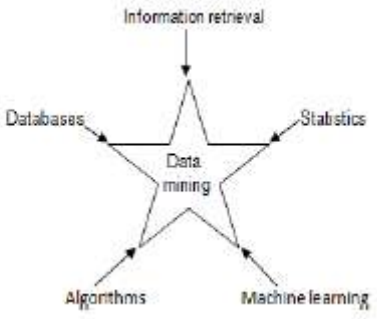

Figure 3.1**:** Historical perspective of data mining

B. History of Database and Data Mining:

Data mining development and the history represented in the figure. The data mining system started from the year of 1960s and earlier. In this, the data mining is simply on file processing. The next stage its Database management Systems to be started year of 1970s early to 1980s. In this OLTP, Data modeling tools and Query processing are worked. From database management system there three broad categories to be worked. First one is Advanced Database Systems, this evaluated year of Mid-1980s to present in this Data models and Application oriented process are worked. The Second part is Data Warehousing and Data Mining worked since the year of the late 1980s to present. The third part is Web based Database Systems this started from 1990s to present and in this Web mining and XML based database systems are included. These three broad categories are joined and create the new process that's called new generation of the Integrated Information system is started in 2000.

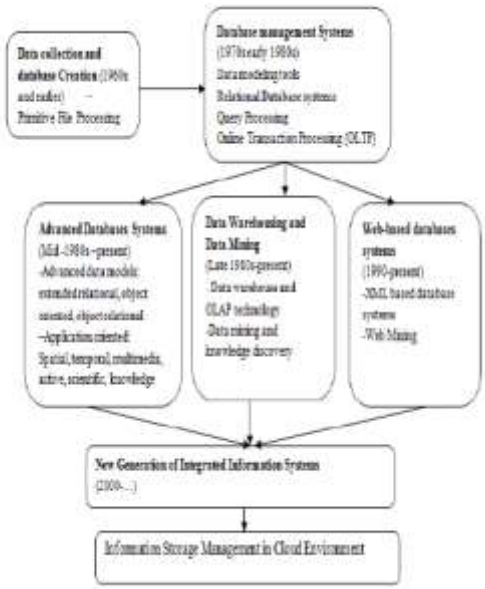

Figure 3.2: History of Database Systems and Data Mining

IV. TYPICAL DATA MINING FRAMEWORK:

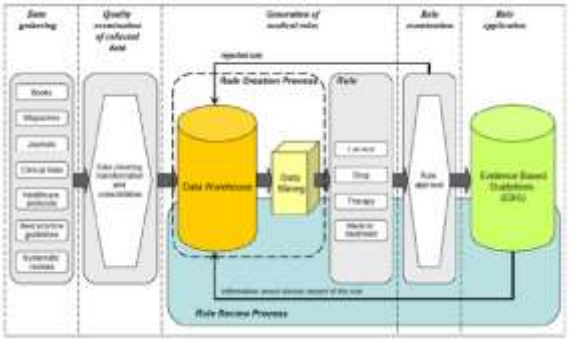

Figure 4.1 Typical Data Mining Framework

Analysis refer to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining for raw data and converting it into information useful for decision making by users. Data is collected and analyzed to answer questions, test hypothesis or disprove theories. There are several phases that can be distinguished, described below. The phases are iterative in that feedback from later phases may result in additional work in earlier phases. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

V. DATA MINING APPLICATIONS IN HEALTHCARE SECTOR

Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into road categories [1,10],

1) Treatment effectiveness

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

2) Healthcare management

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

Using logistic regression models to compare hospital profiles based on risk-adjusted death with 30 days of non-cardiac surgery neural network system to predict the disposition in children presenting to the emergency room with bronchiolitis

Predicting the risk of in-hospital mortality in cancer patients with nonterminal disease

3*) Customer relationship management*

Customer relationship management is a core approach to managing interactions between commercial organizations-typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings. The principles of applying of data mining for customer relationship management in the other industries are also applicable to the healthcare industry. The identification of usage and purchase patterns and the eventual satisfaction can be used to improve overall customer satisfaction. The customers could be patients, pharmacists, physicians or clinics. In many cases prediction of purchasing and usage behavior can help to provide proactive initiatives to reduce the overall cost and increase customer satisfaction.

4) Fraud and abuse

Detect fraud and abuses establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims. Data mining has been used very successfully in aiding the prevention and early detection of medical insurance fraud.

5) Medical Device Industry

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless bio-sensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients.

6) Pharmaceutical Industry

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. A deep understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making.

7) Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized [12]. Three layers of hospital management:

- Services for hospital management
- Services for medical staff
- Services for patients

8) System Biology

Biological databases contain a wide variety of data types, often with rich relational structure. Consequently multi-relational data mining techniques are frequently applied to biological data[13]. Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.

TABLE 5.1: Examples of Research in Data Mining for Healthcare Management:

Researching topic	Researching institute	Dataset
Healthcare data mining: predicting inpatient length of stay	School of Information Management and Engineering, Shanghai University; Harrow School of Computer Science	Geriatric Medicine department of a metropolitan teaching hospital in the UK.
Designing Patient-Specific Seizure Detectors From Multiple Frequency Bands of Intra-cranial EEG Using Support Vector Machines	The Center for Computational Learning Systems (CCLS) and The Columbia University Medical School (CUMC)	Columbia University Medical School has collected approximately 30 TB of intra-cranial EEG recordings.
Classification, Treatment and Management of Alzheimer's Disease Using Various Machine Learning Methods	MGR University, Chennai; UVCE , Bangalore,; Defence Institue of Advanced Technology Pune	National Institute on Aging, USA.

VI. CHALLENGES IN DATA MINING FOR HEALTHCARE

- Missing values, noise, and outliers
- "Cleaning data from noise and outliers and handling missing values, and then finding the right subset of data, prepares them for successful data mining" [Razavi07]
- Transcription and manipulation of patient records often result in a high volume of noise and a high portion of missing values [O'Sullivan08]
- "Missing attribute values can impact the assessment of whether a particular combination of attribute-value pairs is significant within a dataset" [Laxminarayan06]

VII. DATA MINING TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

A. Classification:

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

B. Clustering:

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

C. Predication:

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

D. Association rule:

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

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

E. Neural networks:

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

VIII. WHAT KIND OF DATA CAN BE MINED

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series

databases and textual databases, and even flat files. Here are some examples in more detail:

A. *Flat files:*

Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text 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 measurements, etc.

B. Relational Databases:

Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. In Figure we present some relations Customer, Items, and Borrow representing business activity in a fictitious video store Our Video Store. These relations are just a subset of what could be a database for the video store and is given as an example.

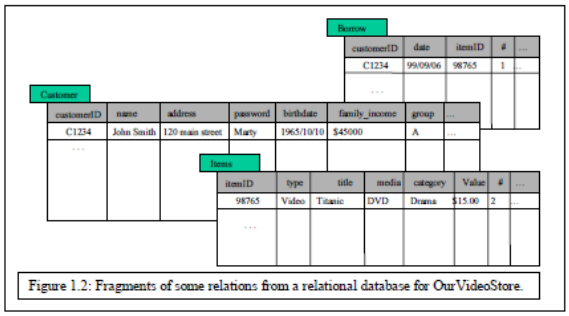

Figure 8.1: Fragment of Some Relation From A Relation Database For Our Video Store

C. Data Warehouses:

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 schema. A data warehouse gives the option to analyze data from different sources under the same roof. Let us suppose that Our Video Store becomes a franchise in North America. Many video stores belonging to Our Video Store company may have different databases and different structures. If the executive of the company wants to access the data from all stores for strategic decision-making, future direction, marketing, etc., it would be more appropriate to store all the data in one site with a homogeneous structure that allows interactive analysis. In other words, data from the different stores would be loaded, cleaned, transformed and integrated together. To facilitate decision making and multi-dimensional views, data warehouses are usually modeled by a multi-dimensional data structure. Figure 1.3 shows an example of a three dimensional subset of a data cube structure used for Our Video Store data warehouse

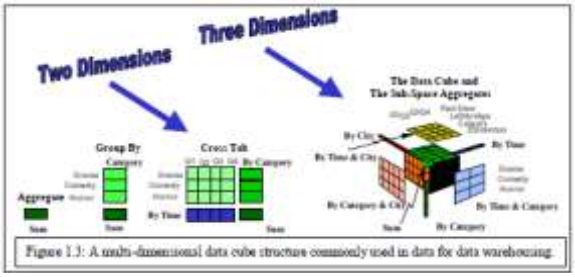

Figure 8.2: A Multi-Dimensional Data Cube Structure Commonly Used In Data For Data Warehouses

The figure shows summarized rentals grouped by film categories, then a cross table of summarized rentals by film categories and time (in quarters). The data cube gives the summarized rentals along three dimensions: category, time, and city. A cube contains cells that store values of some aggregate measures (in this case rental counts), and special cells that store summations along dimensions. Each dimension of the data cube contains a hierarchy of values for one attribute. Because of their structure, the pre-computed summarized data they contain and the hierarchical attribute values of their dimensions, data cubes are well suited for fast interactive querying and analysis of data at different conceptual levels, known as On-Line Analytical Processing (OLAP). OLAP operations allow the navigation of data at different levels of abstraction, such as drill-down, roll-up, slice, dice, etc. Figure 1.4 illustrates the drill-down (on the time dimension) and roll-up (on the location dimension) operations.

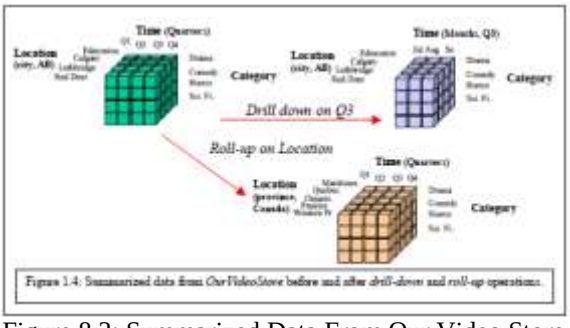

Figure 8.3: Summarized Data From Our Video Store Before And After Drill-Down And Roll-Up Operation

IX. ADVANTAGES OF DATA MINING IN VARIOUS APPLICATONS

Advantages of using data mining in various applications such as Banking, Manufacturing and

production, marketing, health care etc., are as follows[7]:

1) Banking: Data mining supports banking sector in the process of searching a large database to discover previously unknown patterns; automate the process of finding predictive information. Data mining helps to forecast levels of bad loans and fraudulent credit cards use, predicting credit card spending by new customers and predicting the kinds of customer best respond to new loan offered by the backs.

2) Manufacturing and production: Data mining helps to predict the machine failures and finding key factors that control optimization of manufacturing capacity.

3) Marketing: Data mining facilitates marketing sector by classifying customer demographic that can be used to predict which customer will respond to a mailing or buy a particular product and it is very much helpful in growth of business.

4) Health-Care: Data mining supports a lot in health care sector. It supports health care sector by correlating demographics of patients with critical illnesses, developing better insights on symptoms and their causes and learning how to provide proper treatments

5) Insurance: Data mining assist insurance sector in predicting fraudulent claims and medical coverage cost, classifying the important factors that affect medical coverage and predicting the customers' pattern which customer will buy new policies.

6) Law: Law enforcement is helped by data mining by monitoring the behaviour patterns of the criminals. Tracking crime pattern, locations and criminal behaviours, identifying various attributes to data mining, assist in solving criminal cases.

7) Government and Defence: Data mining helps to forecast the cost of moving military equipment and predicting resource consumption. Apart from that it assists in testing strategies for potential military engagements and improving homeland security by mining data from many sources.

8) Brokerage and Securities trading: Data mining assists in predicting the change in bond prices and forecasting the range of stock fluctuation determining when to buy or sell stocks.

9) Computer hardware and software: Predicting disk-failures and potential security violations can be done by data mining.

10) Airlines: It supports in checking the feasibility of adding routes to increase the business profit and to decrease the loss by capturing data on where passengers are flying and the ultimate destination of passengers.

X. X.DISADVANTAGES OF DATA MINING

The disadvantages of data mining are explained as follows:

1) Privacy Issues

One of the disadvantages is a personal privacy issue. In recent years, with the boom of internet, the concerns about privacy have increased tremendously. Because of this privacy concern, individuals like internet users, employees, customers are afraid that unknown person may have access to their personal information and then use that information in an unethical way and this may cause harm to them. Although, several laws have protected the users to sell or trade personal information between different organization, selling personal information have occurred.

2) Security Issues

Another biggest disadvantage is security issue which is always a major concern in information technology. Companies have a lot of personal information about the employees and customers including social security number, birthdates, payroll etc., and it is also available in online. But, they do not have sufficient security systems in place to protect this information. They have been a lot of cases where hackers access and stole personal data of customers.

3) Misuse of Information/Inaccurate information

Trends obtain from the data mining intended to be used for business or some ethical purpose. However it may be misused for other unethical purpose.

XI. FUTURE DIRECTION

Data mining applications in healthcare can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry consider how data can be better captured, stored, prepared, and mined. Possible directions include the standardization of clinical vocabulary and the sharing of across organizations to enhance the bto also explore the use of text mining to expand the scope and nature of what healthcare denefits of healthcare data mining applications. Further, as healthcare data are not limited to just quantitative data, such as physicians'

notes or clinical records, it is necessary ata mining can currently do. In particular, it is useful to be able to integrate data and text mining.36 It is also useful to look into how digital diagnostic images can be brought into healthcare data mining applications. Some progress has been made in these areas.

Figure 10.1: Data Collection and Analysis-e.g. Collecting Data From Social Networking Sites

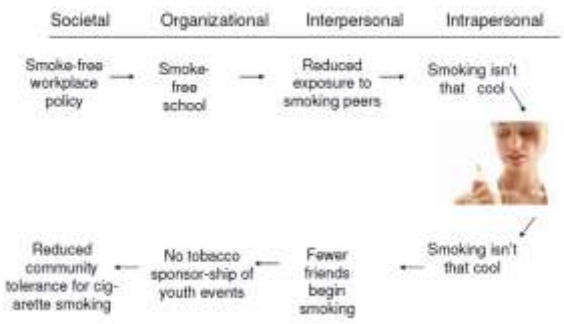

Figure 10.2: Social Ecological Model

XII. CONCLUSION

This paper aimed to compare the different data mining application in the healthcare sector for extracting useful information. The prediction of diseases using Data Mining applications is a challenging task but it drastically reduces the human effort and increases the diagnostic accuracy. Developing efficient data mining tools for an application could reduce the cost and time constraint in terms of human resources and expertise. Exploring knowledge from the medical data is such a risk task as the data found are noisy, irrelevant and massive too. In this scenario, data mining tools come in handy in exploring of knowledge of the medical data and it is quite interesting. It is observed from this study that a combination of more than one data mining techniques than a single technique for diagnosing or predicting diseases in healthcare sector could yield more promising results.

REFERENCES

[1] Prasanna Desikan, Kuo-Wei Hsu, Jaideep Srivastava, "Data Mining For Healthcare Management", 2011 SIAM International Conference On Data Mining, April 28-30, 2011, pp. 14-64.

[2] Mary K. Obenshain, MAT," Application of Data Mining Techniques to Healthcare Data", Statistic For Hospital Epidemiology Vol. 25 No. 8, August 2004, pp. 690-695.

[3] M. Durairaj, V. Ranjani, "Data Mining Applications In Healthcare Sector: A Study", International Journal of Scientific & Technology Research Volume 2, Issue 10, October 2013, pp. 29-32.

[4] Hian Chye Koh and Gerald Tan, "Data Mining Applications in Healthcare", Journal of Healthcare Information Management — Vol. 19, No. 2, June 2005, pp. 64-72.

[5] S.Hameetha Begum, "Data Mining Tools and Trends – An Overview", International Journal of Emerging Research in Management &Technology ISSN: 2278-9359, February 2013, pp. 6-12.

[6] Osmar R. Zaïane, "Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases, 1999, pp. 1-15.

[7] Aqueel Ahmed, Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012, pp. 18-23.

[8] Tijl De Bie, "An Information Theoretic Framework for Data Mining", *KDD'11,* San Diego, California, USA, August 21–24, 2011, pp. 1-9.

[9] Pragnyaban Mishra ,Neelamadhab Padhy, Rasmita Panigrahi, "The Survey of Data Mining Applications And Future Scope", Asian Journal Of Computer Science And Information Technology 2: 4 (2012) , pp. 68– 77.