GNITED MINDS
Journals

International Journal of Information Technology and Management

Vol. V, Issue No. I, August-2013, ISSN 2249-4510

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

A COMPARATIVE ANALYSIS ON USING VARIOUS APPLICATIONS AND TECHNIQUES OF DATA MINING AND DATA WAREHOUSING IN A HEALTHCARE MANAGEMENT

www.ignited.in

International Journal of Information Technology and Management
Vol. V, Issue No. I, August-2013, ISSN 2249-4510

A Comparative Analysis on Using Various Applications and Techniques of Data Mining and Data Warehousing In a Healthcare Management

Vijay S. Jondhale

Research Student, Singhania University, Pacheri Bari, Jhunjhunu, Rajasthan

***Abstract* – Data mining as one of many constituents of health care has been used intensively and extensively in many organizations around the globe as an efficient technique of finding correlations or patterns among dozens of fields in large relational databases to results into more useful health information. In healthcare, data mining is becoming increasingly popular and essential. Data mining applications can greatly benefits all parties involved in health care industry. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform huge amount of data into useful information for decision making.**

This paper illustrates data mining will enable clinicians and managers to find valuable new patterns in data, leading to potential improvement of resource utilization and patient health. As the patterns are based on recent clinical practice, they represent the ultimate in evidence-based care.

Healthcare presents unique challenges for the architect of a data warehouse. Integrated health systems are shifting its focus away from the acute care setting and moving towards cross-continuum care management. Improving healthcare quality while reducing costs requires the elimination of unnecessary variation in the care process. This paper describes the lessons learned during the business case development for the project. Topics include establishing the need for a data warehouse, understanding data warehousing in healthcare, justifying the cost of a data warehouse, building the team, and setting achievable goals.

With continuous advances in technology, increasing number of clinicians are using electronic medical records to accumulate substantial amounts of data about their patients with the associated clinical conditions and treatment details. The 'hidden' relationships and patterns within these information would further our medical knowledge including its efficiencies and deficiencies. Methodologies that are being used in parallel industries with increasing effectivity need to be modified and applied to discover this knowledge. This paper discusses, at a high level, the various methodologies that may be used, along with the elaboration of the various terminologies associated with data warehousing and knowledge discovery in databases (KDD).

The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system.

---------------------------♦-----------------------------

INTRODUCTION

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 massive data stores to uncover previously unknown patterns. Data mining is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Data mining is not new idea, it has been used intensively and extensively by financial institutions for activities such as credit scoring and fraud detection; marketers for direct marketing and cross-selling; retailers for market segmentation and store layout; manufacturers for quality control and maintenance scheduling and it has been used in hospital care as well.Data mining has been becoming increasingly popular, it has been noted that several factors have motivated the use of data mining applications. 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, the application has helped to help them find and track offenders. However fraud detection using data mining applications is prevailing in the commercial world for detection of fraudulent credit card transactions and fraudulent banking activities . Huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods hence calls for technological interventions so as to simplify management of those data. Data mining can improve decision making by discovering patterns and trends in large amounts of complex data. Such analysis has become increasingly essential as financial pressures have amplified the need for healthcare organizations to make decisions based on the analysis of clinical and financial data. Insights gained from data mining can influence cost, revenue and operating efficiency while maintaining a high level of care. Health care organizations that perform data mining are better positioned to meet their long term needs; data can be a great asset to healthcare organizations, but they have to be first transformed into information Yet another factor motivating the use of data mining applications in healthcare is the realization that data mining can generate information that is very useful to all parties involved in the healthcare industry. For example, data mining applications can help healthcare insurers detect fraud and abuse, and healthcare providers can gain assistance in making decisions.. Data mining applications also can benefit healthcare providers such as hospitals, clinics, physicians, and patients by identifying effective treatments and best practices.

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 massive data stores to uncover previously unknown patterns. Data mining is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Data mining is an automated approach for discovering or inferring hidden patterns or knowledge buried in data. ‘Hidden’ means patterns that are not made apparent through casual observation. Data Mining is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms, and statistics. Data Mining is a technology for knowledge discovery in databases (KDD). This technology provides different methodologies for decision making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning and innovation. Data mining is a variety of techniques such as neural networks, decision trees or standard statistical techniques to identify nuggets of information or decision making knowledge in bodies of data, and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting, and estimation.

Electronic medical records are becoming more ubiquitous in day-to-day clinical practice. They capture clinical data, store in personal database as well as mirror it in local and regional databases. Data capture, storage, retrieval and display are all performed. They also allow display of alerts, warnings; guide a clinician through a clinical protocol by way of workflow, and online transactional processing where “intelligent” data display is made through running structured queries using SQL, etc.

Unfortunately, all this data residing in a RDBMS is good enough for basically the following purposes with respect to improving patient care.

1. Display against a period of time allowing for better visualization of the patient’s clinical condition

2. Potentially life-saving alerts/warnings about a patient based on the clinical information collected about the patient They are not able to support either evidence based medicine or outcomes analysis directly. To perform these tasks, it is necessary to have a data warehouse or at least an appropriate custom-built interface for the same. Although running specially designed queries may be able to accomplish this task, the pay off is that the user needs to correctly design them and retrieving results proves to a slow process as the data is usually not “analysis-ready”.

It does however hold the potential to unleash a revolutionizing wealth of information regarding disease processes, disease progression, best method of treatment, optimizing costs while maximizing efficiency, etc. Currently, the way to make this possible is to use online analytical processing by way of using data warehousing and data mining. Data mining, functionally, is the process of discovering interesting knowledge from large amounts of data stored in various data repositories like databases or data warehouses. The process involves integration of techniques likes database technology, statistics, artificial intelligence, high performance computing, data visualization, image/signal processing, and spatial data analysis. By performing this process, interesting knowledge, patterns and high-level information can be extracted, viewed and browsed from multiple angles. The knowledge so discovered can be applied to decision-making, process control, information management, and query processing.

Managing data in healthcare organizations has become a challenge as a result of healthcare managers having considerable differences in objectives, concerns, priorities and constraints. The planning, management and delivery of healthcare services included the manipulation of large amounts of health data and the corresponding technologies have become increasingly embedded in all aspects of healthcare. Information is one of the most factors to an organization success that executive managers or physicians would need to base their decisions on, during decision making.. Healthcare organizations typically deal with large volumes of data containing valuable information about patients, procedures, treatments and etc. These data are stored in operational databases that are not useful for decision makers or executives.

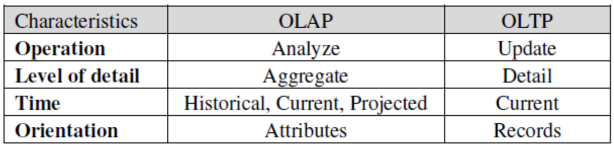
A majority of database management systems used in these organizations execute online transaction processing (OLTP direct answer to queries at the executive level, such as the what-if and what-next type queries. Decision makers at executive level would like to quickly analyze existing health data and in time to aid in the decision making process. However, stand-alone databases cannot provide such information quickly and efficiently. The concept of data warehousing provides a powerful solution for data integration and information access problems.

Data warehousing idea is based on the online analytical processing (OLAP). Basically, this technology supports reorganization, integration and analysis of data that enable users to access information quickly and accurately . The comparison between OLTP and OLAP technology shown in table 1. Finally, OLAP is a tool that used by analyst for planning and decision making.

In traditional information systems, businesses have relied on paper-based reports regarding performance in order to make important business decisions. Most of the reports that are created are out dated; these have come as a result of extracting data from operational systems and collating with other sources of data to come up with them. Managers want and need more information, but analysts can provide only minimal information at a high cost within the desired time frames. A healthcare data warehouse designed for this particular purpose is needed.).

OLTP databases are designed to process individual records of patients, procedures, treatments, drugs and other similar operations. These databases are updated continually and are suitable to support daily operations. Also these databases cannot provide a An important role of a healthcare data warehouse is to provide information for Executive manager to analyze situations and make decisions. Put in another way, a healthcare data warehouse provides information for doctors to make decisions and do their jobs more effectively

The top level of decision making involves strategic decision making. At this level managers make decisions about the overall goals of the organization. For instance, types of decisions made on this level include which services need to be provided (such as acute, ambulatory or long term care) and at which geographical location to operate (such as local, state, national). The second level concerns tactical decision making. The decisions made on this level relate to the tactical units of the organization such as patient care services and marketing. The third level concerns the day to day decisions of the organization such as hiring employees, ordering supplies and medications, processing bills. Good systems provide the information needed, so that Managers are making more efficient decisions. Described in this paper is the development and implement of a prototype healthcare data warehouse specific to cancer diseases, employing the new ‘data warehouse’ technology incorporating large quantity of analysis information needed for healthcare decision-making.



Characteristics	OLAP	OLTP
Operation	Analyze	Update
Level of detail	Aggregate	Detail
Time	Historical, Current, Projected	Current
Orientation	Attributes	Records

Table 1: Comparison of OLAP and OLTP (adapted from Ref)

DATA MINING

Nowadays there is huge amount of data stored in real-world databases and this amount continues to grow fast as it creates both an opportunity and a need for semi-automatic methods that discover the hidden knowledge in such database. If such knowledge discovery activity is successful, discovered knowledge can be used to improve the decision making process of an organization. For instance data about a hospital’s patient might contain interesting knowledge about which kind of patient is more likely to develop a given disease. This knowledge can lead to better diagnosis and treatment for future patients. Data mining and knowledge discovery is the name often used to refer to a very interdisciplinary field, which consists of using methods of several research areas to extract knowledge from real world data sets. There is a distinction between the terms data mining and knowledge discovery; the term data mining refers to the core steps of a broader process called knowledge discovery in database. In addition to the data mining step which actually extracts knowledge from data, the knowledge discovery process includes several preprocessing and post processing steps. The goal of data preparation methods is to transform the data to facilitate the application of a given data mining algorithms, where the goal of knowledge refinement methods is to validate and refine discovered knowledge. The knowledge discovery is both iterative and interactive. It is iterative because the output of each step is often feedback to previous steps and typically many iterations of this process are necessary to extract high quality knowledge from data. It is interactive because the user or more precisely an expert in the application domain should be involved in this loop to help in data preparation, discovered-knowledge validation and refinement.

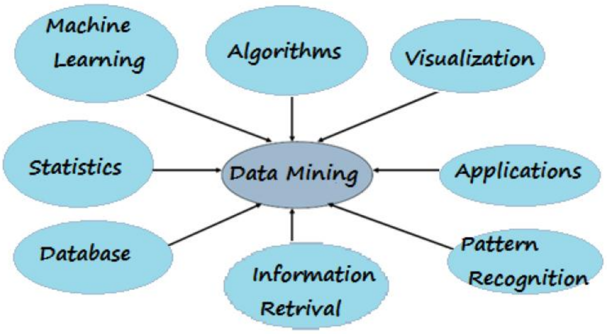
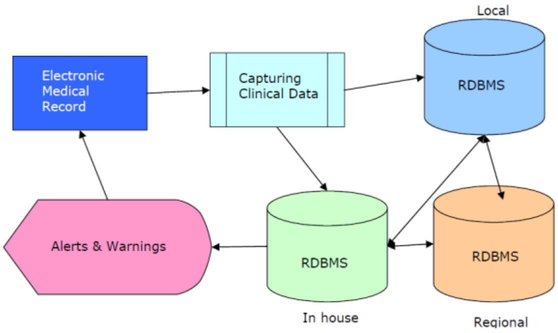


Fig 1: Data Mining Architecture.

Current Status

Electronic medical records capable of capturing clinical data, storing them locally (i.e. the clinician's own database), in-house (i.e., in the same organization like clinic, department or hospital), and regionally (i.e. in the same geographical area), capable of displaying data on request, alerts that are rule-based and patient-specific (display an alert if this patient's systolic blood pressure comes down below 60 mmHg or fasting blood sugar is more than 110 mg/dL on three consecutive days, etc.), warnings that have been pre-set (display a warning if any patient allergic to penicillin group of drugs and is suffering from rheumatic fever with ASO titer more than 200 Todd units, or a contra-indicated drug is prescribed, or two interacting drugs are prescribed concomitantly, etc.), following clinical protocols and performing other OLTP functions. The relevant software architecture is as follows.



DATAMINING TECHNIQUES IN HEALTH CARE

Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective, 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. Other data mining applications related to treatments include associating the various side-effects of treatment, collating common symptoms to aid diagnosis, determining the most effective drug compounds for treating sub-populations that respond differently from the mainstream population to certain drugs and determining proactive steps that can reduce the risk of affliction. Future needs of individuals to improve their level of satisfaction. These applications also can be used to predict other products that a healthcare customer is likely to purchase, whether a patient is likely to comply with prescribed treatment or whether preventive care is likely to produce a significant reduction in future utilization.

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining. We now describe a few Classification data mining techniques with illustrations of their applications to healthcare.

A. Rule set classifiers :

Complex decision trees can be difficult to understand, for instance because information about one class is usually distributed throughout the tree. C4.5 introduced an alternative formalism consisting of a list of rules of the form "if A and B and C and ... then class X", where rules for each class are grouped together. A case is classified by finding the first rule whose conditions are satisfied by the case; if no rule is satisfied, the case is assigned to a default class IF conditions.

THEN conclusion - This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction.

B. Decision Tree algorithms:

Decision tree include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node. CART uses Gini index to measure the impurity of a partition or set of training tuples . It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient have a certain disease or not.

C. Neural Network Architecture :

The architecture of the neural network used in this study is the multilayered feed-forward network architecture with 20 input nodes, 10 hidden nodes, and 10 output nodes. The number of input nodes is determined by the finalized data; the number of hidden nodes is determined through trial and error and the number of output nodes is represented as a range showing the disease classification. The most widely used neural-network learning method is the BP algorithm.

Learning in a neural network involves modifying the weights and biases of the network in order to minimize a cost function. The cost function always includes an error term a measure of how close the network's predictions are to the class labels for the examples in the training set.

Additionally, it may include a complexity term that reacts to a prior distribution over the values that the parameters can take. Neural networks have been proposed as useful tools in decision making in a variety of medical applications. Neural networks will never replace human experts but they can help in screening and can be used by experts to double-check their diagnosis. In general, results of disease classification or prediction task are true only with a certain probability.

D. Neuro-Fuzzy :

Stochastic back propagation algorithm is used for the construction of fuzzy based neural network. The steps involved in the algorithm are as follows: First, initialize weights of the connections with random values. Second for each unit compute net input value, output value and error rate. Third, to handle uncertainty for each node, certainty measure (c) for each node is calculated. Based on the certainty measure the decision is made.

USING A DATA WAREHOUSE in healthcare

The concept of "data warehousing" arose in mid 1980s with the intention to support huge information analysis and management reporting. Data warehouse was defined According to Bill Inmon a “subject-oriented, integrated, time variant and non-volatile collection of data in support of management’s decision making process” .

According to Ralph Kimball “a data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making".

Today, data warehouses are not only deployed extensively in banking and finance, consumer goods and retail distribution and demand-based manufacturing, it has also became a hot topic in noncommercial sector, mainly in medical fields, government, military services, education and research community etc.

A data warehouse is typically a read-only dedicated database system created by integrating data from multiple databases and other information sources. A data warehouse is separate from the organization’s transactional databases (i.e., OLTP databases). It differs from transaction systems in that :

- It covers a much longer time horizon (several years to decades) than do transaction systems.

- It includes multiple databases that have been processed so that the warehouse’s data are subject oriented and defined uniformly (i.e., ‘‘clean prearranged data’’).

- It contains non-volatile data (i.e., read-only data) which are updated in planned periodic cycles, not frequently.

- It is optimized for answering complex queries from direct users (decision makers) and applications.

Data warehouse architecture is a description of the components of the warehouse, with details showing how the components will fit together. Data is imported from several sources and transformed within a staging area before it is integrated and stored in the production data warehouse for further analysis. Since a data warehouse is used for decision making, it is important that the data extracted from multiple sources should be corrected. It is inevitable that when different data are integrated into the data warehouse, there is a high probability of errors and anomalies. Therefore, tools for data extraction, data cleaning, data integration and finally data load are required. Data are stored and managed in the warehouse which presents multidimensional views of data to a variety of front end tools: query tools, report writers, analysis tools, and data mining tools. The architecture of data warehousing is shown in Figure 2:

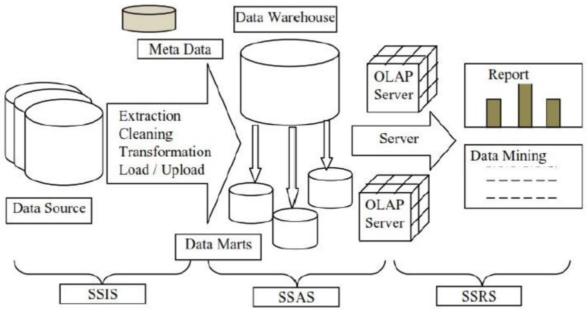


Figure2. Data Warehouse Architecture

Decision support applications based on operational transaction systems already existed within individual business units of the IDS. Management justifiably asked why existing systems were not sufficient. To answer this question, an elucidation of the risks, benefits and costs associated with a data warehouse were required. First, we considered OUT risks:

- Few people understood the differences between operational and decision support systems.

- It was difficult to conceptualize the added benefit of data integration.

- The target user groups were not accustomed to an interactive data interface since the standard had been static reporting.

- There was a concern that it would be difficult to sustain support throughout the life of the project.

METHODOLOGY

Steps of Knowledge Discovery

- Cleaning to remove data inconsistencies and aberrations (called “noise”)

- Extraction of data from multiple sources and integrating them into a data warehouse (it is a common practice to perform data cleaning and data integration as a pre-processing step before storing the resultant data)

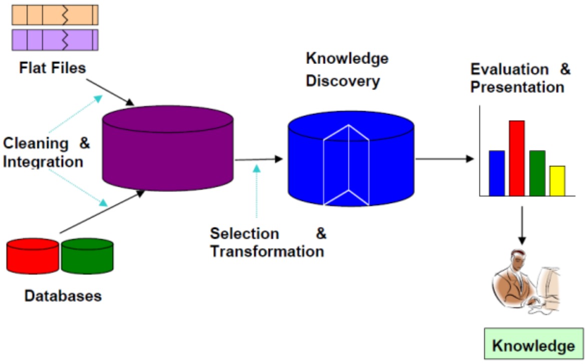
- Selection of data relevant to the analysis task by retrieving them from the data base

- Transformation of data that are consolidated into forms appropriate for mining by performing summary or aggregation operations (occasionally data transformation and consolidation are performed before the data selection process, particularly in data warehouses)

- Data mining is the essential process where intelligent methods are applied in order to extract data patterns

- Evaluation of extracted data pattern is performed to identify the truly interesting patterns representing knowledge

- Knowledge presentation is carried out by using visualization and knowledge representation techniques that are used to present the mined knowledge to the user The process is pictorially depicted below.



The process of getting data out of databases and into data warehouses is no easy task. This is the first step and the most time-consuming one. The next is to create, where necessary, data marts. This needs to be followed by data mining through framing appropriate queries and running them. The results display assumes vital importance for clinicians since only when a particular data is presented in a particular way does it become meaningful – a series of values is less preferable to a graphical display, while the value of density of a tissue with the image would convey more meaning than the image alone.

All the data clinical captured through electronic medical records can provide invaluable insights into the trends, progression, patterns and management of disease and its processes through the process of knowledge discovery using data analysis techniques.

The entity-relationship data model is commonly used in the design of relational databases (RDBMS), where a database schema consist of a set of entities and the relationships between them. This is usually a two dimensional in nature.

Such a data model is appropriate for OLTP. On the other hand, a data warehouse requires a concise, subject-oriented schema that facilitates OLAP. The most popular data model for a data warehouse is a multidimensional model (three dimensional or cubic, where the each data item has three or more contexts associated with it). Just as relational query languages like SQL is used to specify relational queries, a data mining query language (DMQL) is used to specify data mining tasks.

CONCLUSION

The use of data mining has focused on evidence-based patterns from previous patient treatment. In all likelihood, the absence of automated discovery of patterns would leave many questions unasked. These questions, if asked, would benefit not only the resource utilization for patient treatment, but also the health of the patient.

Data mining helps professionals discover these patterns and put them to work. As models are based directly on history, they represent the ultimate in evidence-based care. But technology is no panacea, and professional, ethical and practical issues must be addressed. Decisions must rest with the healthcare professionals, not the information systems experts.

The concept, usefulness and practicality of data warehousing is already a fact. They are being tested and rolled out in increasing numbers world wide. The return-on-investments have well justified the costs involved in building and maintaining one. It is but a natural progression on the path of information and knowledge management.

In a clinical setting, the first step is to capture, authenticate, validate, store and retrieve the data in the proper manner. This is already being done through electronic medical records. The next is to unearth the knowledge that lies “hidden” within the captured data.

This is accomplished through clinical data warehousing and data mining using a team of health analysts. This team is made up of medical specialists and data mining technology experts. While the specialists help by framing the “right” questions for analysis, the technologists do the actual data model designing and the tasks of ETL and data mining. The results are passed back to the specialists who then perform the task of discovery and report the findings to the concerned persons.

We studied the problem of constraining and summarizing different algorithms of data mining. We focused on using different algorithms for predicting combinations of several target attributes. In this paper, we have presented an intelligent and effective heart attack prediction methods using data mining. Firstly, we have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weightage, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Five mining goals are defined based on business intelligence and data exploration. The goals are to be evaluated against the trained models.

In our future work, this can further enhanced and expanded. For predicting heart attack significantly 15 attributes are listed. Besides the 15 listed in medical literature we can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. We can also use Text Mining to mine the vast amount of unstructured data available in healthcare databases.

REFERENCES

- A.A. Freitas, "Understanding the crucial role of attribute interaction in data mining," Artificial Intelligence Review, vol. 16, pp. 177-199, 2001.

- Craig S. Ledbetter, Matthew W. Morgan. Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse

- Dale Sanders. Healthcare Analytics: Standing on the Brink of a Revolution. Journal of Healthcare Information Management, Vol. 16, no. 4

- Gilbreath, R E. Health care data repositories: components and a model. The Journal of the Healthcare Information and Management Systems Society, Vol. 9, No. 1, 63-73.

- J. Dych (2000) “E-Data Turning Data into Information with Data Warehousing”, Addison-Wesley, Reading.

- Jiawei Han, Micheline Kamber. Data mining – concepts and techniques. Morgan Kaufmann Publishers. ISBN – 1055860-489-8

- Kolar, H.R. (2001). Caring for healthcare. Health Management Technology, 22(4), 46-47.

- Laura Hadley, (2002) “Developing a Data Warehouse Architecture”.

- N. Padhy, "THE SURVEY OF DATA MINING APPLICATIONS AND FEATURE SCOPE Pragnyaban Mishra, Neelamadhab Padhy," 2012.

- Niederman, F. (1997). Data warehousing at an urban hospital. Jml of Data Warehousing, Vol. 2, No. 4, ~~2-12.

- Philip Baylis et al. Better health care with data mining, Clementine – working with health care. SPSS white paper. Shared Medical Systems Limited, UK

- Silver, M. Sakata, T. Su, H.C. Herman, C. Dolins, S.B. & O’Shea, M.J. (2001). Case study: how to apply data mining techniques in a healthcare data warehouse. Journal of Healthcare Information Management, 15(2), 155-164.

- T.-H. Chen and C.-W. Chen, "Application of data mining to the spatial heterogeneity of foreclosed mortgages," Expert Systems with Applications, vol. 37, pp. 993-997, 2010.

- V. Poe, P. Klauer, S. Brobst (1998) “Building a Data warehouse for Decision Support”, Prentice Hall, Upper Saddle River.