Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 3, March 2017, pg.160 – 168

Data Mining & Warehousing Algorithms and its Application in Medical Science - A Survey

Sumitha Thankachan[1], Suchithra[2]
[1]Assistant Professor, Department of Computer Science, BVM Holy Cross College, Cherpunkal, India
[2]Graduate Trainee, Academic Coordinator Office, Amrita VishwaVidyapeetham(ASE), Coimbatore, India
sumitha.thankachan@yahoo.com, chanu.suchithra@gmail.com

Abstract: One among the fastest growing fields is health care industry. The medical industry contains large amount of medical data which would not be "mined". The mined data helps in finding the hidden information. Extensive amount of data in medical database need the development of tools which are used to access the data, analyze the data, knowledge discovery, and efficient use of the stored knowledge and information. The medical industry have large amount of data collected about the patient including the details, diagnosis and medications. Turning these data into useful pattern helps in predicting with the new treatments and medicines. This helps in the better diagnosis and therapy where the patients can attain the good QoS (quality of service). This paper features the different data mining and warehousing techniques used in healthcare field for the best decision making.

Keywords - Knowledge Discovery, Medical Database.

I. INTRODUCTION

The main purpose of data mining is for the extraction of the useful and relevant information from the large databases or data warehouses. Applications of Data mining are mainly useful for commercial and scientific areas [1]. This study discusses mainly on the Data Mining applications in the scientific area. Data mining in scientific area distinguishes itself in the sense that the nature of the datasets are often very different from the traditional market driven applications of data mining. A

detailed survey is done on data mining applications in this work on healthcare sector, the types of data used and the details of the information extracted as output. Data mining algorithms which are applied in the healthcare industry plays a significant role in the prediction and the diagnosis of the diseases. There are a huge number of data mining applications that are found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management.

Therefore, to find the useful and hidden knowledge from the database is the main purpose behind the application of data mining. Data mining is also called as knowledge discovery from the data. As the name itself suggests, knowledge discovery is an interactive and iterative process, which consists of developing and understanding the application domain, selection and creation of a data set, preprocessing and data transformation.

In health care institutions, data mining tools answer the question rapidly, that are traditionally a time consuming and too complex to resolve. They prepare the databases to find the predictive information.

The Expanding of the health coverage to many people as possible and to provide financial assistance to help them with the lower income purchase coverage [2]. To Eliminate the health disparities that are in the current situation, it is better to decrease the costs that are associated with the increased disease burden borne by certain population growth. Healthcare administration is a field which is related to the leadership, management, and administration of hospitals, hospital networks, and health care systems[1,3].

Healthcare sector focuses mainly on:

- The proposal of the draft of NHP 2001 which is timely to the State health expenditure which is to be raised up to 7% by 2018 and 8% of State budgets thereafter [21].
- Health spending in India at 6% of GDP is among the highest levels estimated for developing countries.
- Public spending on health in India has declined itself after the liberalization from 1.3% of GDP in 1990 to 0.9% in 1999. Central budget allocations for health have stagnated at 1.3% of the total Central budget. In the States it has declined from 7% to 5.5% of the State health budget.

This paper focuses on the comparison of the data mining tools with the health care problems. The comparative study helps in finding the accuracy level to be predicted by the data mining applications in the healthcare. This comparative study leads the aspiring researchers in the field of data mining by knowing which data mining tool gives an accuracy level in extracting information from healthcare data.

Data Mining had been used in a variety of applicative areas such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile and mobile computing.

Some data mining applications are:

- To detect fraudulent phone or credit-card activity
- Predicting good and poor sales
- Predicting cardiovascular disease.
- To detect defects in manufacturing process.

II. LITERATURE REVIEW

A literature review is conducted on the applicability of data mining in medical field which has the critical points of the current knowledge also including the substantial, theoretical and methodological contributions.

The **[1]** paper mainly discusses on the data mining and its applications including the major areas of the treatment effectiveness, Managing the healthcare, also the detection of the fraud and also provides an overview about the customer relationship management.

The **[2]** paper presents how data mining helps in discovering and also in extracting the useful patterns of the large data to find the possible observable patterns. This paper encompasses the importance and the ability of Data mining in improving the quality of the decision making process in the medical industry.

[3] Illustrates a combination of the prediction system which includes Rough Set Theory (RST) and Artificial Neural Network (ANN) for the dispensation of the medical data. The process of developing a new data mining technique and a software for assisting the competent solutions for medical data analysis has been explained. This paper also proposes a hybrid tool which incorporates RST and ANN for making a proficient data analysis and also indicative predictions. The experiments on data set for the prediction of excellence of animal semen is carried out. The projected system is applied for pre-processing of a medical database and also to train the ANN for the production of prediction. The predicted accuracy is observed for the comparison of the observed and predicted cleavage rate [20].

[4] discusses mainly on the potential use of the classification which is based on data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of medical data. The parameters that are considered here are age, sex, blood pressure and blood sugar which can predict the likelihood of patients getting a heart disease.

[5] Has discussed about the various data mining approaches which have been utilized for the diagnosis of breast cancer and for the prognosis of the Decision tree is found as the best predictor with the utmost accuracy of 93.62%.

[6] Has discussed about the disease caused by HIV that weakens the body that can no longer fight the simple infections. The algorithm is used to discover association rules. WEKA 3.6 is used to mine the data to implement the algorithms, J48 classifier performs the classification with an accuracy rate of 81.8%.

[7] Discussed what a Data mining can contribute to the blood bank sector. The algorithm used here is J48 algorithm and the tool used is WEKA. Classification rules performs well in the classification with

an accuracy rate of 89.9%.

Apriori **[18]** algorithm is used for the frequent item set mining and also for the association rule learning over the transactional databases. It also proceeds with the identification of the frequent individual items in the database and extending them to the larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets are determined by the Apriori which can be used to determine the association rules which usually highlight the general trends in the database which also includes the applications in domains such as market basket analysis.

Apriori uses breadth-first search and a Hash tree structure for counting the candidate item sets efficiently. It also generates candidate item sets of same length. It then prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent-length item sets. After that, it also scans the transaction database to determine frequent item sets among the candidates. The pseudo code for the algorithm is given below for a transaction database, and a support threshold of .Usual set theoretic notation is employed, though note that is a multi-set is the candidate set for level . At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. Accesses a field of the data structure that represents candidate set, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.

$$
\begin{array}{l}
\textbf{Apriori}(T, \epsilon) \\
\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
\quad k \leftarrow 2 \\
\quad \textbf{while } L_{k-1} \neq \emptyset \\
\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\} \\
\quad\quad \textbf{for transactions } t \in T \\
\quad\quad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
\quad\quad\quad \textbf{for candidates } c \in C_t \\
\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
\quad\quad L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\} \\
\quad\quad k \leftarrow k + 1 \\
\quad \textbf{return } \bigcup_k L_k
\end{array}
$$

III. DATA MINING

A process in which a raw data is being prepared and structured such that valuable information can be extracted from it is called Data analysis. The process of organizing and thinking about data is way to accepting what the data does and does not contain. Data Analysis is a process of inspecting, cleaning, transforming, and modeling data. The objective of data analysis is to highlight useful information, providing conclusions, and help in decision making. Data analysis consists of multiple steps and

approaches, including diverse techniques under an array of names, in different business, science, and social science domains. [9]

The data mining process is an automatic or semi-automatic analysis of huge amount of data for the extraction of interesting patterns of data records known as cluster analysis, a group of unusual records for anomaly detection, and to find out dependencies i.e., association rule mining and sequential pattern mining. The usual database techniques are spatial indices. These patterns are used in further analysis i.e., in machine learning and predictive analytics.

Data Mining is the discovery of unknown information from the databases [15] [20]. Data Dredging, data fishing and data snooping refer to the use of data mining method to sample part of a larger population data set which are too small for reliable statistical inferences to be made to validate the patterns discovered. These methods can be used in the creation of new hypothesis to test data against the larger data.

Data mining functions:

- Clustering,
- Classification,
- Prediction, and
- Associations.

Currently the evaluation of data mining functions and products are the results of the influence from many of the disciplines, which includes the databases, information retrieval, statistics, algorithms, and machine learning [9] (See Fig. 1).

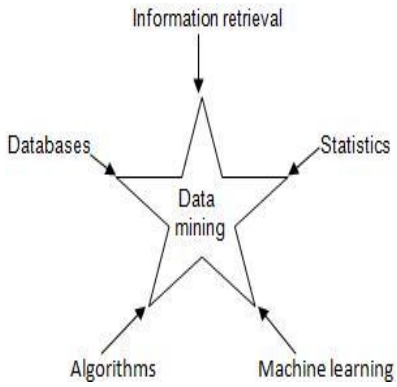

Fig. 1. Historical perspective of data mining

3.1 Data Base and Data Mining – A Review

The development of Data mining is represented in the Fig. 2. The system of data mining started in early 1960s. Here data mining is just a file processing. The next stage of it is Database management Systems which started in the year 1970s and was still under process till early 1980s. Here OLTP, Data modeling tools and Query processing worked.

There are three broad categories in which a database management system worked.

- First is Advanced Database Systems, which was evaluated in the Mid-1980s to present. In this Data models and Application oriented process worked.
- The Second part is Data Warehousing and Data Mining which worked since late 1980s to present.
- The third part is Web based Database Systems which started from 1990s to present. This includes Web mining and XML based database systems.

These are the three broad categories are joined and created a new process called Integrated New generation Information system which was started in 2000.

3.2 Data Mining Application Areas

Data Mining is driven by new applications, which requires new inclusions that are not currently in use.

These are classified into two categories:

Business & E-Commerce.

Scientific, Engineering & Healthcare Data.

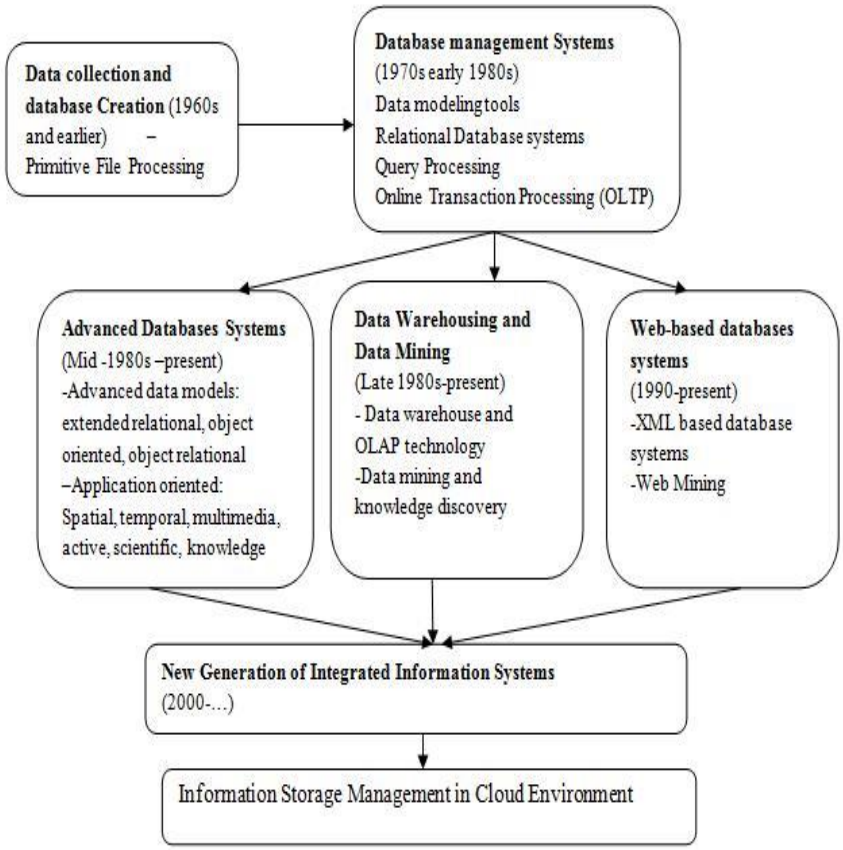

Fig. 2. History of Database Systems and Data Mining

3.3 Data Mining Tasks

Data mining tasks are mainly classified into two broad categories:

- Predictive model
- Descriptive model

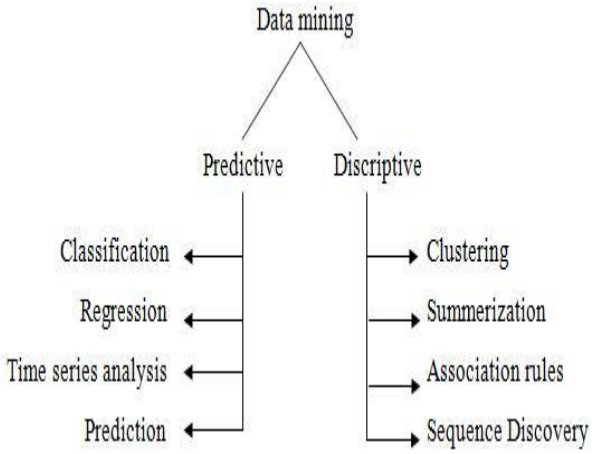

Fig 3.3 Data mining models and tasks

IV. DATA MINING APPLICATIONS IN HEALTHCARE SECTOR

Healthcare industry generates huge amounts of data about patients, resources, diagnosis, electronic patient records, medical devices etc. Larger amounts of data are the key resource which are to be processed and analyzed for knowledge extraction which enables the support for cost-savings and decision making.

Data mining applications in healthcare can be grouped as the evaluation into broad categories [1,10],

4.1 Treatment effectiveness

Data mining applications are developed for the evaluation of the effectiveness of medical treatments. Data mining can deliver analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

4.2 Pharmaceutical Industry

The technology used here 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 positional and organizational decision-making.

4.3 Hospital Management

Modern hospitals are capable of generating and collecting huge amount of data. Mined data are stored in a hospital information system where a temporal behavior of global hospital activities to be visualized [12].

Three layers of hospital management:

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

4.4 System Biology

Other Database which has huge amount of data is Biological databases that contain a wide variety of data types, usually with rich relational structures. Subsequently multi-relational data mining techniques are frequently applied in biological data [13].

V. CONCLUSION AND FUTURE WORK

In this paper, a study of how data mining techniques are used for the data analysis and Knowledge discovery in medical sciences is carried out. This paper aimed only for the comparison of the different data mining applications in the healthcare sector for extracting useful information. It is a challenging task, the prediction of diseases using Data Mining applications but it drastically reduces the human efforts and also increases the diagnostic accuracy. A future work may be implemented in developing efficient data mining tools for an application could reduce the cost and time constraint in terms of human resources and expertise.

REFERENCES

1. HianChyeKoh and Gerald Tan, Applications in Healthcare Information Management –Vol 19, No 2.
2. JayanthiRanjan, ―Applicati techniques in pharmaceuticJournalof Theoretical and Applied Technology, (2007).
3. RubanD.Canlas Jr., MSIT., Healthcare: Current applications.
4. K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ― Applications of Data Mining Healthcare and Prediction - International Journal on Computer Science and Engineering (2010).
5. ShwetaKharya,singData Mining―U Techniques ForDiagnosis And Prognosis - International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.
6. EliasLemuye,tatus Predictive―Hiv Modelings using Data Mining Technology Ⅱ.
7. Arvind Sharma and P.C. Guhan - Number of Blood Donors through their Age and Blood Group by using Data Mining techniques - International Journal of Communication and Computer Technologies Volume 01 –No.6, Issue: 02 September 2012.
8. Arun K PunjDatari,Mining― Tec Universities (India) Press Private Limited, 2006.
9. Boros E., P.L. Hammer, T. Ibaraki, A. Kogan.(1997). Logical Analysis of Numerical Data. Mathematical Programming, 79:163-190.
10. Boros E., P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik.(2000). An Implementation of Logical Analysis of Data. IEEE Transactions on knowledge and Data Engineering, 12(2):292-306.
11. Crama Y., P.L. Hammer, T. Ibaraki. (1988). Cause-effect Relationships and Partially Defined Boolean Functions. Annals of Operations Research, 16:299-325.
12. David Wai-Lok Cheung, Vincent T. Ng, Ada Wai-Chee Fu, and Yongjian Fu. (December 1996). Efficient Mining of Association Rules in Distributed Databases, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 911-922.
13. E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz and I. Muchnik(December 1996) An implementation of logica analysis of data, RUTCOR Research Report RRR 22-96, Rutgers University, 1996., pp. 911-922.
14. Hammer P.L.(1986). The Logic of Cause-effect Relationships, Lecture at the International Conference on Multi-Attribute Decision Making via Operations Research-based Expert Systems, Passau, Germany.
15. Hannu Toivonen (1996). Sampling Large Databases for Association Rules, Proceedings of the 22nd International Conference on Very Large Databases, pp. 134-145, Mumbai, India.
16. Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo (July 1994). Efficient Algorithms for Discovering Association Rules, Proceedings of the AAAI Workshop on Knowledge Discovery in Databases (KDD-94), pp. 181-192.
17. H.J.Adèr,. (2008). Chapter 14: Phases and initial steps in data analysis. In H.J. Adèr & G.J. Mellenbergh (Eds.) (with contributions by D.J. Hand), Advising on Research Methods: A consultant's companion (pp. 333-356). Huizen, the Netherlands: Johannes van Kessel Publishing.
18. http://www2.cs.uregina.ca/~dbd/cs831/notes/itemsets/item set prog1.htm
19. Irshad Ullah, Abdus Salam and Saif-ur-Rehman (2008),Dissimilarity Based Mining for Finding Frequent itemsets.

Proceedings of 4th international conference on statistical sciences Volume (15), University of Gujrat Pakistan, 15: 78
20. Irshad Ullah (2010). Data Analysis by Data Mining Algorithms A Practical Apprach. International Conference on Word Statistics Day. Superior University Lahore
21. Jiawe Han & Micheline Kamber, Data mining concepts and techniques San Franciso Moraga Kaufman 2001.
22. M. Houtsmal and A. Swami (1995). Set-Oriented Mining for Association Rules in Relational Databases, Proceedings of the 11th IEEE International Conference on Data Engineering, pp. 25-34, Taipei, Taiwan.
23. Ming-Syan Chen, Jiawei Han and Philip S. Yu.(1996). Data Mining: An Overview from a Database Perspective, IEEE transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866-883
24. Peter L. Hammer Tiberius Bonates.(2005). Logical Analysis of Data: From Combinatorial Optimization to Medical Applications, RUTCOR Research Report RRR 10 - 2005

