An Empirical Study of the Applications of Web Mining Techniques in Health Care

Dr. Varun Kumar
Department of Computer Science & Engineering
ITM University,
Gurgaon, India

MD. Ezaz Ahmed
Department of Computer Science & Engineering
ITM University,
Gurgaon, India

***Abstract*-Few years ago, the information flow in health care field was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. There has already been a long standing tradition for computer-based decision support, dealing with complex problems in medicine such as diagnosing disease, managerial decisions and assisting in the prescription of appropriate treatment. Today, one of the biggest challenges that health care system, face is the explosive growth of data, use this data to improve the quality of managerial decisions. Web mining and Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This paper addresses the applications of web mining and data mining in health care management system to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples. Further we propose the IDSS model for the health care so that exact and accurate decision can be taken for the removal of a particular disease.**

Keywords- Web mining; Health care management system; Data mining; Knowledge discovery; Classification; Association rules; Prediction; Outlier analysis, IDSS.

I. Introduction

In modern world a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in the field of Medical science, Agriculture, Business and so on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database management tools are no longer adequate for analyzing this huge amount of data. Particular disease in a particular area and it will be related with the agriculture of the particular area so that we can predict prone of particular disease in a particular area due to excessive use of pesticides and fertilizer in a particular agricultural area. We have to collect clinical data for a particular disease and with the help of data mining and web mining we can predict a pattern.

Web mining as well as Data Mining (sometimes called data or knowledge discovery) has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information. There are increasing research interests in using web mining and data mining in health care or web based health care management. This new emerging field, called health care Web mining, concerns with developing methods that discover knowledge from data come from medical environments [1].

The data can be collected from various medical institutes, doctor's clinic or hospital that resides in their databases. The data can be personal or institutional which can be used to understand patients' behavior, to assist doctor, to improve diagnosis, to evaluate and improve health care systems , to improve error free or zero error treatment and many other benefits.[1][2]

Health care data mining used many techniques such as decision trees, neural networks, k-nearest neighbor, naive bayes, support vector machines and many others. Using these methods many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for organization of syllabus, to predict how many patients will register for a particular disease, alienating traditional clinical model predicting a particular disease in a particular area.

This paper is organized as follows: Section 1 introduction and describes the data mining techniques adopted. Section 2 discusses the application areas of these techniques in a medical institute or clinic or any hospital. Section 3 concludes the paper. Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. [5]

A. Data mining techniques

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The steps identified in extracting knowledge from data are:

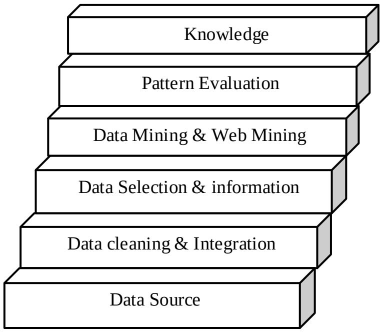


Figure1. The steps of extracting knowledge from data

B. Association analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

More formally, association rules are of the form $X \Rightarrow Y, i.e., "A_i \wedge ---- \wedge A_m \rightarrow B_j \wedge ---- \wedge B_n"$, where Ai (for i to m) and Bj (j to n) are attribute-value pairs. The association rule X=>Y is interpreted as database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y ".

C. Classification and Prediction

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction.

IF-THEN rules are specified as IF condition THEN conclusion

e.g. IF age=old and patient=diabetic then heart disease prone=yes

D. Clustering Analysis

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity.

That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. [5]Application of clustering in medical can help medical institutes' group individual patient into classes of similar behavior. Partition the patient into clusters, so those patients within a cluster (e.g. healthy) are similar to each other while dissimilar to patient in other clusters (e.g. disease prone or Weak).

E. Outlier Analysis

A database may contain data objects that do not comply with the general behavior of the data and are called outliers. The analysis of these outliers may help in disease detection and predicting abnormal values or reason of disease.

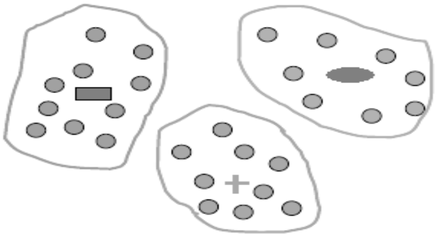


Figure 2 Picture showing the partition of patients in clusters

II. POTENTIAL APPLICATIONS

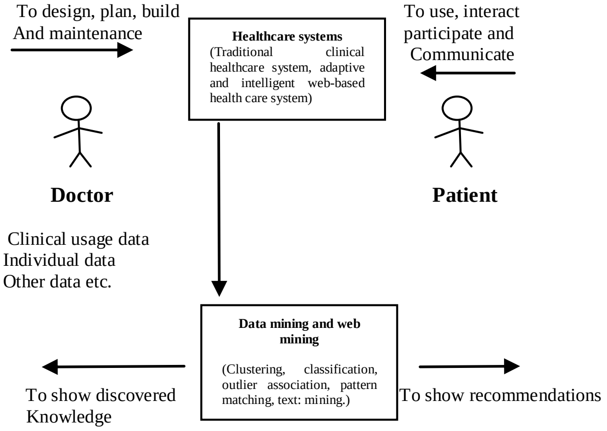


Figure 3 The cycle of applying web mining or data mining in health care system

The figure No. 3 illustrates how the data from the traditional clinic and web based health care systems can be used to extract knowledge by applying web mining and data mining techniques which further helps the doctors and patients to make decisions.

A. Organization of DATA

It's important for medical institutes to maintain a high quality healthcare program. This will improve the patient's cure process. This will also use for the optimization of resources.

Presently, complication of disease is influenced by many factors such as life style, other disease, environment fertilizer, pesticides, genetic history, availability of best doctor, expert team of doctors and experiences.

One of the applications of data mining is to identify related disease in period of institutional programmes in a large healthcare institute.

A case study has been performed where the patient data collected over a period of time at healthcare Institute. The main of the study was to find the strongly related disease in a period offered by the institute. For this purpose following methodology was followed:

1) *Identify the possible related disease.*
2) *Determine the strength of their relationships and determine strongly related disease.*

In the first step, association rule mining was used to identify possibly related two disease combinations in the period which also reduces our search space. In the second step, Pearson Correlation Coefficient was applied to determine the strength of the relationships of disease combinations identified in the first step. [4]

TABLE 1 SHOWS THE PATIENT SUFFERING FROM THE DISEASE

Patient id	Disease 1	Disease 2	Disease 3
1	Blood Sugar	Blood Pressure	Heart Disease
2	Blood Sugar	Blood Pressure	Heart Disease
3	Blood Sugar	Blood Pressure	Heart Disease
4	Blood Sugar	Blood Pressure	Kidney Problem
5	Blood Sugar	Blood Pressure	Eye Problem

Association Rules that can be derived from Table 1 are of the form:

$(X, disease1) \Rightarrow (X, disease2)$

$(X, disease1)^\wedge(X, disease2) \Rightarrow (X, disease3)$

$(X, "Bloodsugar") \Rightarrow (X, "Bloodpressure")$ [support=2% and confidence=60%]

$(X, "Bloodsugar")^\wedge(X, "bloodpressure" \Rightarrow (X, "Heartdisease")$ [support=1% and confidence=50%]

Where support factor of the association rule shows that 1% of the patient suffering from the disease blood sugar and blood pressure, confidence factor shows that there is a chance that 50% of the patients who have “Blood sugar” will also have “Blood pressure”

This way we can find the strongly related disease and can optimize the database of a healthcare programme.

B. Predicting the Registration of Patients in an Hospital

Now healthcare organizations are getting strong competition from itself, the need of better health by the government plans and by the people concern. It needs deep and enough knowledge for a better assessment, diagnosis, planning, and decision making.

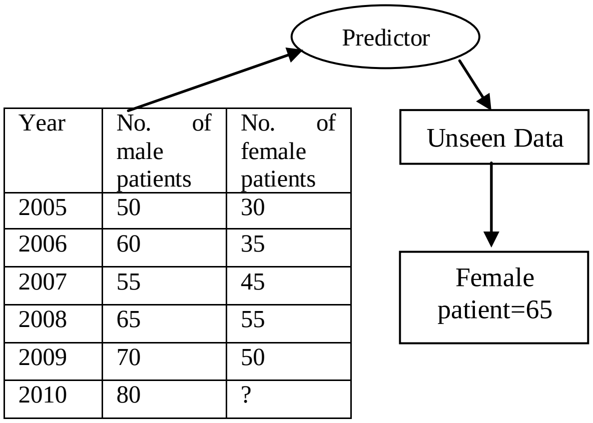
Data Mining helps hospitals to identify the hidden patterns in databases; the extracted patterns are then used to build data mining models, and hence can be used to predict diagnosis and decision making with high accuracy. As a result of this hospitals are able to allocate resources more effectively. [6]

One of the applications of data mining can be for example, a hospital can take necessary actions before patient quit their treatment, or to efficiently assign resources with an accurate estimate of how many male or female will register for a particular disease by using the Prediction techniques.

In real scenario couple of other associated attributes like type of disease, hygiene environment, climate etc. can be used to predict the registration of patients in a particular disease.

C. Predicting Patients Response Against Medicine

One of the questions, whose answer almost every doctor or patient of a hospital would like to know “Can we predict patient performance against the medicine?” [6]



Year	No. of male patients	No. of female patients
2005	50	30
2006	60	35
2007	55	45
2008	65	55
2009	70	50
2010	80	?

Figure 4 Prediction of female patient in the coming year

Over the years, many researchers applied various data mining techniques to answer this question.

In modern times, healthcare system is taking on a more important role in the development of our civilization by providing good health. Good health maintenance is an individual behavior as well as a social phenomenon.

It is a difficult task to deeply investigate and successfully develop models for evaluating healthcare system efforts with the combination of modernization of equipment’s and practice. Healthcare organization’s goals and outcomes clearly relate to “promoting good health through effective healthcare models and research in service to healthcare system.” With the help of data mining techniques a decision making system can be developed which can help doctors and patients to know the weak points of the traditional clinical healthcare model. Also it will help them to face the rapidly developing real-life environment and adapt the current healthcare realities (such as IDSS etc.)

We use patients’ response to medicine participation data as part of the grading policy. A doctor can assess the condition of patient by conducting an online discussion among a group of patients and use the possible indicators such as the time difference between medicine taken and time of response to the medicine etc.[6] With the help of this data, we can apply classification algorithms to classify the patients into possible levels of grading.

D. Identifying Abnormal/ Erroneous Values

The data stored in a database may reflect outliers-|noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. [5]

One of the applications of Outlier Analysis can be to detect the abnormal values in the response sheet of the patient. This may be due many factors like a software fault, data entry operator negligence or an extraordinary response of the patient for a particular disease.

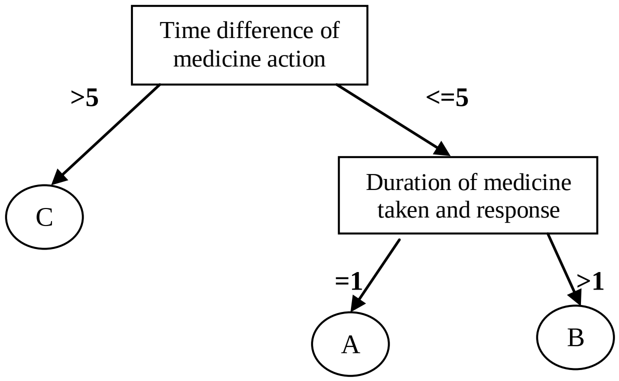

Time difference of medicine action
>5
<=5
C
Duration of medicine taken and response
=1
>1
A
B

Figure 5 The Decision Tree built from the data in Table 2

TABLE 2 PATIENT RESPONSES DATA AND THEIR GRADES

Time difference between medicine taken action (in min)	Duration between medicine taken and response (in min)	Grade of the Patients
3	1	A
3	2	B
4	1	A
5	2	B
6	1	C
6	2	C

An outlier is an observation that is numerically distant from the rest of the data. Outliers, being the most extreme observations, may include the sample maximum or sample minimum or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

TABLE 3 THE RESPONSE OF PATIENT IN FOUR DISEASES

Patient ID	Level Blood Sugar		ECG Reading	Blood pressure reading	Heart stroke chance
	FF	PP			
101	130	231	35	145/89	30
102	167	270	75	128/90	67
103	189	310	90	178/100	77
104	230	450	35	190/105	**99**

In the table shown above the response of the patient in disease 4 heart strokes with patient ID 104 will be detected as an exceptional case and can be further analyzed for the cause.

III. CONCLUSION

The focus of the study was to discuss the various data mining techniques which can support healthcare system via generating strategic information.

Since the application of data mining brings a lot of advantages in higher well equipped hospitals, it is recommended to apply these techniques in the areas like optimization of resources, prediction of disease of a patient in the hospital, the disease response of the patient, number of them respond, number of the cured and number of them which are fully satisfied.

REFERENCES

[1] C. Romero, S. Ventura, E. Garcia, "Datamining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008

[2] C. Romero, S. Ventura "Educational dataMining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007

[3] Shaeela Ayesha, Tasleem Mustafa, AhsanRazaSattar, M. Inayat Khan, **"**Data Mining Model for Higher Education System", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010

[4] W. A Sandham, E.D. Lehman "SIMULATING AND PREDICTING BLOOD GLUCOSE LEVEL FOR IMPROVED DIABETES HEALTHCARE", cmru,Imperical College, Royal Brompton London

[5] Han Jiawei, MichelineKamber, *Data Mining: Concepts and Technique.* Morgan Kaufmann Publishers,2000

[6] A. Hasman, R. Bindels and P. de Clercq "On the use of reminder Systems in Healthcare", Department of Medical Informatics, University of Maastricht, Netherlands. IEEE International Conference

[7] Sri Lanka Institute of Information Technology, *http://www.sliit.lk/*

[8] Sun Hongjie, "Research on Student Learning Result System based on Data Mining", IJCSNS International Journal of Computer Science and Network Security, Vol.10, No. 4, April 2010

[9] Academy Connection – Training Resources Inhtml, http:*//www.cisco.com/web/learning/netacad/index,* December 28th, 2005.

[10] Wayne Smith, "Applying Data Mining to Scheduling Courses at a University", Communications of the Association for Information Systems, Vol. 16, Article 23.

[11] Kumar, V. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. IJACSA - International Journal of Advanced Computer Science and Applications, 2(3), 80-84.

[12] Thakur, M. (2011). Query based Personalization in Semantic Web Mining. IJACSA - International Journal of Advanced Computer Science and Applications, 2(2), 117-123.

AUTHOR'S PROFILE

Prof. Varun Kumar

Ph.D. (Computer Science), Associate Head , CSE Deptt. School of Engineering and Technology, ITM University, Gurgaon Haryana ,India. Presently 3 Ph. D students are working under his supervision. Dr. Varun Kumar, completed his PhD in Computer Science. He received his M. Phil. in Computer Science and M. Tech. in Information Technology. He has 13 years of teaching experience. He is recipient of Gold Medal at his Master's degree. His area of interest includes Data Warehousing, Data Mining, and Object Oriented Languages like C++, JAVA, C# etc. He has published more than 35 research papers in Journals/Conferences/Seminars at international/national levels. He is working as an Editorial Board Member / Reviewer of various International Journals and Conferences. He has 3 books, 5 study materials and 3 lab manuals to his credit.

Md. Ezaz Ahmed

Pursuing Ph.D. under the supervision of Prof, Varun kumar. Currently, he is working with itm University as Asst. Professor. He did his M.E (CSE) in first division with honors. Before joining this Institute, He has more than 17 years of experience out of which 15 years teaching and 2 years industry experience. He has published 11 research papers, 1 in international Journal others in national conference and in departmental journal. His area of interest includes Web Development, software Engineering, Software verification validation and testing, Soft Computing, and Basics of computer and C programming. He is a member of Indian Society of Technical Education (ISTE). Co-author of one project book published in 1998.