THE IMPACT OF DATA MINING TECHNIQUES ON MEDICAL DIAGNOSTICS

***Siri Krishan Wasan*[1]*, Vasudha Bhatnagar*[2] *and Harleen Kaur*[1*]**

[*1]*Department of Mathematics, Jamia Millia Islamia, New Delhi, India.*
Email: harleen_k1@rediffmail.com
[2] *Department of Computer Science, University of Delhi, New Delhi, India.*

ABSTRACT

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. Data mining and statistics both strive towards discovering patterns and structures in data. Statistics deals with heterogeneous numbers only, whereas data mining deals with heterogeneous fields. We identify a few areas of healthcare where these techniques can be applied to healthcare databases for knowledge discovery. In this paper we briefly examine the impact of data mining techniques, including artificial neural networks, on medical diagnostics.

Keywords: Data mining, Knowledge discovery in databases (KDD), Healthcare administrators, Health data, Medical diagnostics

1 INTRODUCTION

It is well known that in an Information Technology (IT) driven society, knowledge is the most significant asset of any organization. The role of Information Technology in health care is well established. Continuing advancement in IT applications in the area of healthcare has raised people's expectations for better and less expensive healthcare. Computerization of hospital information systems (HIS) has provided an easy access to valuable medical information. A hospital information system not only covers billing, a patient's admission, payroll, budget, and accounting but also can improve the speed and quality of work in relation to laboratories, pharmacies, and medical documentation.

A comprehensive hospital information system to meet the specific needs of a hospital contains modules for in-patient/out-patient registration, patient care, pharmacy, diet planning, accounting, etc. Sophisticated equipment used in the practice of modern medicine generates huge amount of data. This data is usually stored in digital form, and considerable effort is being made to find automated methods of data analysis to generate knowledge. This knowledge can be used for fast and better clinical decision-making.

The area of data mining and knowledge discovery tools can play an effective role in achieving these goals (Cios & Moore, 2000). Knowledge discovery is a well-defined process consisting of several steps including data mining. Recently, data mining has invoked significant interest in its applications to healthcare (Cios & Moore, 2000). The healthcare environment is usually information rich but knowledge poor. However, data mining techniques can be applied to create a knowledge rich healthcare environment. For example, applications of data mining techniques to Acquired Immune Deficiency Syndrome (AIDS) datasets can be a highly important area. In a thickly populated country with scarce resources such as India, information dissemination and knowledge discovery from large databases seem to be the only solution to check the spread of AIDS.

In view of the large amount of medical data being generated, there is growing pressure for improved methods of data analysis and knowledge discovery using appropriate data mining techniques. A proper medical database created with intention mining can provide a useful resource for data mining and knowledge discovery. The terms knowledge discovery in databases (KDD) and data mining are used interchangeably. KDD is the process of finding useful information, and data mining is the process for extracting knowledge (information and patterns) derived by the KDD process using algorithms etc.

2 CHARACTERISTICS OF MEDICAL DATA

Many packages have been developed for hospital information systems, and some hospitals have developed their own systems. There are several necessary application areas for a specific hospital information system. Medical data include data about a patient's history, laboratory reports, in-patient data, out-patient data, nursing staff information, specialists' information, other employees' records, surgery schedule, pharmacy information, billing, budget, etc.

The traditional paper medical record is not only cumbersome but hides much important medical information. However, computerized medical records, containing information about a patient's history, physical examination details, and laboratory findings are more legible and can be made available at several points in the hospital. Computerization of medical records requires some departure from the usual record keeping approach. Moreover, the coding schemes need to be more flexible to allow different doctors to record patient information with different levels of specificity.

Standard unified coding schemes may not work in a medical information system. In general, a system may include the following application areas:

- Outpatient registration
- Emergency records
- Inpatient reservation/ registration
- Patient history
- Patient care
- Pharmacy
- Laboratory reports
- Consultants details
- Employees data
- Account data
- Billing

Medical data are different from data in other databases. Medical data are heterogeneous and involve text and images to a great extent. For example, MRI, ECG or EEG, and cardiac SPECT generate huge amounts of heterogeneous medical data. Knowledge discovery from this data can greatly benefit mankind in the form of improved diagnostic techniques.

Medical data are characterized by their heterogeneity with respect to data type. These data may be noisy with erroneous or missing values. The records of millions of patients can be stored and computerized. However, there are other important issues such as ownership and privacy related to these records. For example, cancer epidemiology is an important area of medical science where anatomic pathology reports can generate huge amount of data to be mined for epidemiologic distribution of cancer (Cios & Moore, 2000). A pathology report is a written medical document which describes the analysis of specimens by the pathologist. These specimens are sent to the laboratory which stores the information. Questions exist about the ownership of the data – are they the property of the patient, the laboratory, or the pathologist?

There is a need to develop methods for mining different types of these data including X-ray, MRI images, electrocardiogram ECG signals, and cholesterol level. The medical history of a patient is an important factor in determining the nature of treatment for the patient. We may consider the medical records with respect to a particular disease D as points in multidimensional data space that is:

$$D = \{(v_1, v_2 \ldots v_n) \mid v_i \in D_i\}$$

where D_i are domain values of various attributes A_i of the disease D. We may introduce the concept of distance between records in dataspace. For example, if records are taken as binary n-tuples (i.e. attributes take values 0 and 1), then the distance can be taken as the Hamming distance defined as follows:

are two medical records, where $x_1, x_2, x_3 \ldots x_n$ and $y_1, y_2, y_3 \ldots y_n$ are attributes (features) of two patients $\underline{u}$ and $\underline{v}$ respectively, then the distance between $\underline{u}$ and $\underline{v}$ is defined as $d(\underline{u}, \underline{v})$ = number of places where $\underline{u}$ and $\underline{v}$ differ.

It is easy to check that:

If $\underline{u} = (x_1, x_2, x_3 \ldots x_n)$ and $\underline{v} = (y_1, y_2, y_3 \ldots y_n)$

(i) $d(\underline{u}, \underline{v}) \geq 0$ and $= 0$ if $\underline{u} = \underline{v}$

(ii) $d(\underline{u}, \underline{v}) = d(\underline{v}, \underline{u})$

(iii) $d(\underline{u}, \underline{v}) \leq d(\underline{v}, \underline{w}) + d(\underline{w}, \underline{v})$

We may use coding techniques to encode data sets into a codeword by adding parity bits to ensure the correctness of the data. The distances between data points can also be defined as Cartesian distances. Once medical records are represented as points in a multidimensional data space, we can identify clusters by visual inspection, which can help in making predictions. Visualization techniques are useful methods for discovering patterns. For example, heart patients may have interesting patterns with respect to levels of blood sugar. Once interesting subsets are obtained, we may use other data mining techniques to discover further knowledge.

Raza Abidi (2001) has emphasized the involvement of knowledge management in the healthcare enterprise. Abidi contends that the operational efficiency of a healthcare enterprise can be increased by using empirical knowledge to drive a suite of packaged strategic healthcare decision support services (SHDS) derived from healthcare data and health enterprise knowledge bases. Specific types of SHDS include analysis of trends of admissions, treatments patterns, forecasting new diseases to evolve appropriate preventive measures, and forecasting complications during treatments.

3 DATA MINING TECHNIQUES IN MEDICAL DATABASES

The large growth of medical databases in advanced countries has motivated medical researchers to use data mining for knowledge discovery from these databases. As the volume of stored data increases, data mining techniques assume an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. Data mining techniques can help answer several critical questions, such as:

- Given the records of dialysis patients, what can be done to improve the treatment of these patients?
- Given the historical patient records on cancer, should the treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?
- Can human DNA databases be characterized as genetic coding models?

Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. Without data mining, it is difficult to realize the full potential of data collected within an organization, as the data is massive, highly dimensional, distributed, and uncertain. In many medical diagnostics, data mining techniques can be used for knowledge discovery. Potential uses of data mining techniques, including artificial neural networks for medical diagnostics, have been effectively demonstrated by (Scales & Embrechts, 2002) and (Kraft, Desouza, & Androwich, 2003).

Data mining is not a single technique. It relates to the idea that there is more knowledge hidden in the data than what they show. Data mining techniques form a group of heterogeneous tools and techniques and are used for different purposes. These techniques and methods are based on statistical techniques, visualization, machine learning, etc. In the first instance, we can apply simple Structured Query Language (SQL) to a dataset to obtain much information but not the hidden knowledge. One can start with some simple statistical techniques such as averages. For example,

- What is the average age of persons suffering from diabetes?
- What is the average age of persons suffering from heart disease?
- What is the average period of survival after angioplasty/ heart operation/ cancer treatment?
- What is the average age of persons admitted to a hospital suffering from a heart problem / cancer/ diabetes?
- What is the average hemoglobin of women of a particular community?

It is interesting to see how these averages change when we focus on different diseases or when we focus only on males or only on females or persons belonging to a particular area or particular regions. 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.

Data mining involves using different algorithms to accomplish different tasks. Basically, the algorithms try to fit a model closest to the characteristics of data under consideration. Models can be predictive or descriptive. Predictive models are used to make predictions, for example, to make a diagnosis of a particular disease. A patient may be subjected to particular treatment not because of his own history but because of results of treatment of other patients with similar symptoms. Descriptive models are used to identify patterns in data. Classification, regression, and time series analysis are some of the tasks of predictive modeling, whereas clustering, association rules, and visualization are some are the tasks of descriptive modeling. We briefly describe some of the basic data mining tasks:

Classification maps or classifies a data item into one of several predefined classes. A set of classification rules is generated from the classification model, based on the features of the data in the training set, which can be used to classify future data and develop a better understanding of each class in the database. For example, classification rules about diseases can be extracted from known cases and used for diagnosis of new patients based on their symptoms.

This is the most important data mining technique, and medical diagnosis is an important application of classification. We may classify patients with heart problems on the basis of various types of heart diseases. Some knowledge of data under consideration is assumed before applying the classification technique.

Suppose D is a database of patients. We may regard D as set of tuples (x_1, x2 $_{...}$ x_n) where x_1, x2 $_{...}$ x_n are values of attributes A_1,A_2 $_{...}$ A_n relevant to a particular disease. We may define various classes C= {C_1, C2 $_{...}$ C_n} of patients depending on severity of disease or particular classification type of the disease. The classification problem is basically to define a function; $f = D \rightarrow C$ where each $t_i \in$ D is mapped to $f\,(t_i)$ belonging to some C_j.

Regression is a method to map target data using some known type of function. It deals with estimation of an output value based on input values.

Time series analysis is the value of an attribute examined over a time period usually at evenly spaced time intervals. For example, depending upon the conditions of a patient, values of certain attributes may be obtained on a daily or hourly basis. This may be used to predict future values or to determine similarity between different time intervals.

Predictive data modeling is an important data mining task to determine future data states on the basis of past and current values. Predictions may be made on the basis of regression, time series analysis, or some other approaches.

Visualization techniques are useful methods of discovering patterns in a medical data set. Scatter diagrams in a Cartesian plane of two interesting medical attributes can be used to identify interesting subsets of medical data sets. For example, for heart patients interesting subsets can be found with respect to blood sugar (fasting). Once interesting subsets are obtained, we may use other data mining techniques on these subsets to discover further knowledge.

An association rule is the discovery of associations among objects. An association rule is in the form of "A_1 ^ A_2 …^ A_i => B_1 ^ B_2^…^ B_j" which means objects B_1 ^ B_2^….^ B_j (B_1 and B_2 …and B_j) tend to appear with objects A_1 ^ A_2 …^ A_i (A_1 and A_2 … and A_i) in the target data. For example, one may discover that a set of symptoms often occur together with another set of symptoms.

Clustering is the identification of classes or clusters for a set of unclassified objects based on their attributes. It is a knowledge discovery process to find groups of interrelated cases and the statistical behaviors that make them adhere into groups. Once the clusters are decided, the objects are labeled with their corresponding clusters, and common features of the objects in a cluster are summarized to form the class description. For example, a set of new diseases can be grouped into several categories based on the similarities in their symptoms, and the common symptoms of the diseases in a category can be used to describe that group of diseases.

4 ARTIFICIAL NEURAL NETWORK – A DIAGNOSTIC TOOL

Neural Networks are analytical techniques modeled after the (hypothesized) process of learning in the cognitive system and the neurological functions of the brain. They are capable of predicting new observations from other observations after executing a process called learning from existing data. Neural networks or artificial neural networks (ANN) are also called connectionist systems, parallel distributed systems, or adaptive systems because they are composed of a series of interconnected processing elements that operate in parallel. A neural network can be defined as a computational system consisting of a set of highly interconnected processing elements, called neurons, which process information as a response to external stimuli. Stimuli are transmitted from one processing element to another via synapses or interconnection, which can be excitatory or inhibitory. Neural networks are good for clustering, sequencing, and predicting patterns.

Neural Networks are one of many data mining analytical tools that can be utilized to make predictions about key healthcare indicators such as cost or facility utilization. Neural networks are known to produce highly accurate results in practical applications. Scales & Embrechts (2002) have used artificial neural networks and fuzzy logic to build a model for heart disease prediction. After performing the usual data cleaning and data separation, they applied principal component analysis (PCA) and extracted features that serve as the best regression predictors (Scales & Embrechts, 2002). Further, they employed partial least square regression (PLS) to discover additional prediction features as shown in Figure 1. Scales and Embrechts found that a non-linear feature selection method such as ANN serves as a better predictor for heart disease. In the preliminary stage of knowledge discovery, they utilized linear data mining techniques. These techniques provided initiation points for artificial neural networks and fuzzy logic analysis as shown in Figure 1. They used several techniques to facilitate the creation of a heart disease diagnostic system. Using cardiovascular disease datasets from the University of California Irvine (UCI) Data Repository, they divided heart disease into five classes though their method can be applied to any serious disease D (Blake & Merz, 2001). For example, we may divide the various stages of the seriousness of the disease.

Using domain knowledge (expert knowledge), one can identify various attributes x_i significant for a particular disease D. Then through preliminary discovery by statistical methods and capturing knowledge from experts, one can assign initial connection weights w_i to the variables x_i respectively. A typical neural network may have several hidden layers, and each layer can have several neurons. We may choose attributes (features) in such a way that layers are connected using a feed forward network, i.e. connections travel in one direction throughout the ANN. (Figure 2)

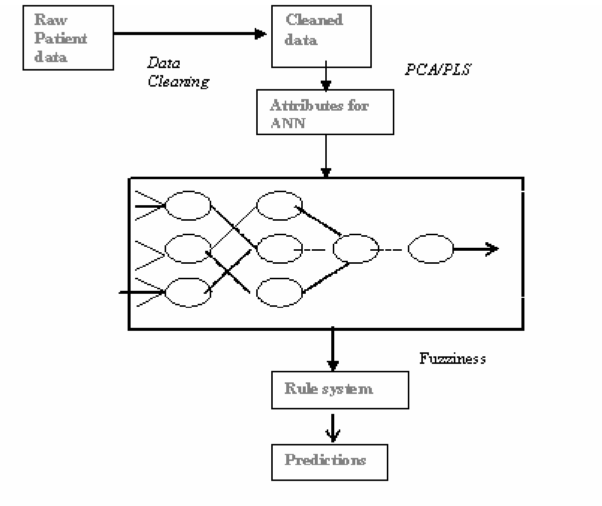

Figure 1. Neural networks with outcome predictions

As the ANN is a machine learning algorithm, the system can learn new knowledge by adjusting connection weights. The neural network architecture and algorithm characteristics determine learning ability. With the adjustment of connection weights between layers, we can achieve better performance. Using data mining techniques such as neural networks, we can transform domain knowledge to create a predictive model.

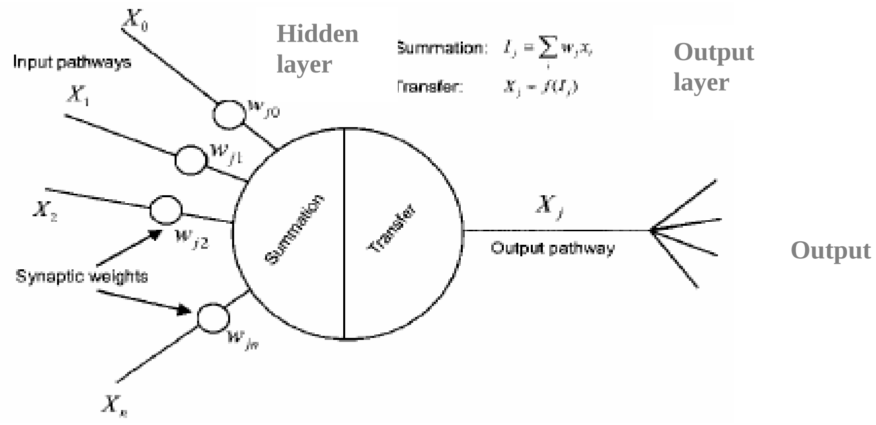

Figure 2. Showing the neurons and its functions

Artificial neural networks provide a powerful tool to help doctors analyze, model, and make sense of complex clinical data across a broad range of medical applications. In medicine, neural networks have been used to analyze blood and urine samples, track glucose levels in diabetics, determine ion levels in body fluids, and detect pathological conditions such as tuberculosis (Lundin, 1998). Neural networks have been successfully applied to various areas of medicine, such as diagnostic aides, medicine, biochemical analysis, image analysis, and drug development**.** They are used in the analysis of medical images from a variety of imaging modalities. Applications in this area include tumor detection in ultra-sonograms, detection and classification of microcalcifications in mammograms, classification of chest x-rays, and tissue and vessel classification in magnetic resonance images (MRI). At the Pacific Northwest National Laboratory [http://www.emsl.pnl.gov:2080/proj/neuron/neural/neural.homepage.html], neural networks are being developed to examine thallium scintigram images of the heart and identify the existence of infarctions (Itchhaporia et al., 1996). Another project at Pacific Northwest National Laboratory used neural network technology to aid in the visualization of three-dimensional ultrasonic images (Miller, 1993). The National Institutes of Health has used artificial neural networks as tools in the development of drugs for treating cancer and AIDS (Miller, Blott, & Hames, 1992). Neural networks are also used in the process of modeling biomolecules. Neural networks may be used in diagnosing ailments such as heart murmur, coronary artery disease, lung disease, and epilepsy.

Neural networks have been applied within the medical domain for clinical diagnosis, image analysis and interpretation (Miller, Blott & Hames, 1992; Miller, 1993), signal analysis and interpretation, and drug development (Weinstein & Kohn, 1992). Neural networks can be used to extract rules from a disease classification. From the rules system so discovered, we can predict if someone will have a particular stage of a particular disease D.

5 SUMMARY

Many legal issues are associated with any use of medical databases, but with proper permission of appropriate authority and adequate care with respect to the confidentially of patient data, we may discover significant useful knowledge. Clinical data when collected contains many errors and therefore needs to be standardized and checked for accuracy and reliability. Computers are no doubt faster than humans in performing mathematical calculations, but human brains can perform many complex tasks such as image and speech recognition in a better way. For example, the artificial neural network, an important data mining technique, attempts to capture, to some extent, this aspect of brainpower in computer models.

Medical records are highly sensitive and contain a large amount of personal information. No doubt, traditional record keeping methods are also vulnerable regarding privacy and security, but remote access to a patient's data raises further risks of unauthorized entry into the data system. Thus employees connected with computerized data management system should be subjected to careful screening before employment. The creation of a proper audit trail by the database administrator will act as a deterrent to the misuse of medical information.

6 ACKNOWLEDGEMENT

We are grateful to Dr. Mrs. Manju Kanga, Associate Specialist, Wycompe Hospital, London, U.K. for her suggestions and valuable discussions.

7 REFERENCES

Cios, K. J., & Moore, G. W. (2000) Medical Data Mining and Knowledge Discovery: An Overview. In Cios K. J., *Medical Data Mining and Knowledge Discovery.* Heidelberg: Physica–Verlag.

Scales, R., & Embrechts, M. (2002) Computational Intelligence Techniques for Medical Diagnostics. Proceedings of Walter Lincoln Hawkins, Graduate Research Conference 2002 from the World Wide Web: http://www.cs.rpi.edu/~bivenj/MRC/proceedings/papers/researchpaper.pdf

Blake, C., & Merz, C.J. (2001) UCI Repository of Machine Learning Databases. [Machine-readable data repository]. University of California, Department of Information and Computer Information and Computer Science, Irvine, C.A. [Available fromhttp://www.ics.uci.edu/~mlearn/MLRepository.html]

Lundin, J. (1998) Artificial Neural Networks in outcome prediction. *Anns Chir Gynaecol* 87, 128-130.

Itchhaporia, D., Snow P.B., Almassy, R. J., & Oetgen, W. J. (1996) Artificial Neural Networks: current status in cardiovascular medicine. *J Am Coll Cardiol.* Aug, 28(2), 515-521.

Miller, A., Blott, B., & Hames, T. (1992) Review of Neural Network Applications in Medical Imaging and Signal Processing. *Medical and Biological Engineering and Computing* 30(5), 449-464.

Miller, A. (1993) *The Application of Neural Networks to Imaging and Signal Processing in Astronomy and Medicine*, PhD thesis, Faculty of Science, Department of Physics, University of Southampton, U.K..

Weinstein, J.N., Kohn, K.W., et al. (1992) Neural Computing in Cancer Drug Development: Predicting Mechanisms of Action. *Science* (258), 447-451.

Kraft, M.R., Desouza, K.C., & Androwich, I. (2003) Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population, *Proceedings 36th Hawaii International Conference on System Sciences (HICSS'03).*

Abidi, S.S.R. (2001) Knowledge management in healthcare: towards 'knowledge-driven' decision –support services. *International Journal of Medical Informatics* 63, 5-18.