Journal of Computer Science 2 (2): 194-200, 2006
ISSN 1549-3636

Empirical Study on Applications of Data Mining Techniques in Healthcare

Harleen Kaur and Siri Krishan Wasan
Department of Mathematics, Jamia Millia Islamia, New Delhi-110 025, India

Abstract: 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. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, decision tree and Artificial Neural Network to massive volume of healthcare data. In particular we consider a case study using classification techniques on a medical data set of diabetic patients.

Key words: Healthcare, health data, medical diagnosis, data mining, artificial neural network, knowledge discovery in databases (KDD).

INTRODUCTION

It is well known that in Information Technology (IT) driven society, knowledge is one of the most significant assets of any organization. The role of IT in health care is well established. Knowledge Management in Health care offers many challenges in creation, dissemination and preservation of health care knowledge using advanced technologies. Pragmatic use of Database systems, Data Warehousing and Knowledge Management technologies can contribute a lot to decision support systems in health care.

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows:

"Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data"[1].

Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Following are some of the important areas of interests where data mining techniques can be of tremendous use in health care management.

* Data modeling for health care applications
* Executive Information System for health care
* Forecasting treatment costs and demand of resources
* Anticipating patient's future behavior given their history
* Public Health Informatics
* e-governance structures in health care
* Health Insurance

Traditionally, decision making in health care is based on the ground information, lessons learnt in the past resources and funds constraints. However, data mining techniques and knowledge management technology can be applied to create knowledge rich health care environment.

A health care organization may implement Knowledge Discovery in databases (KDD) with the help of a skilled employee who has good understanding of health care industry. KDD can be effective at working with large volume of data to determine meaningful pattern and to develop strategic solutions. Health care analyst and policy makers can learn lessons from the use of KDD in other industries and apply KDD to problems of health care industry (Hospitals, Insurance companies, Physicians and Pharmaceutical companies etc.).

Health care data is massive. It includes patient centric data, resource management data and transformed data. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care. Following are a few examples of such questions.

Corresponding Author: Miss. Harleen Kaur, Research Scholar, Department of Mathematics, Jamia Millia Islamia, New Delhi-110 025, India, Tel: +91-9891174111

- Should the course of treatment for a cancer patient include Chemotherapy alone or Chemotherapy plus radiation or radiation alone?
- What can a doctor do to improve his efficiency for treating a dialysis patient?
- Can human DNA databases be sampled up against diseases to produce genetic coding models?
- Can suspicious billing fraudulent be checked?

However there can be a concern of patient privacy. It is more than clear that the role of data mining is not to practice medicine but to improve useful information and knowledge so that better treatment and health care be provided.

Knowledge discovery in medical databases: Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry[2,10]. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven. Data mining is a young interdisciplinary field closely connected to data warehousing, statistics, machine learning, neural networks and inductive logic programming.

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 healthcare organization as data under analysis is massive, highly dimensional, distributed and uncertain.

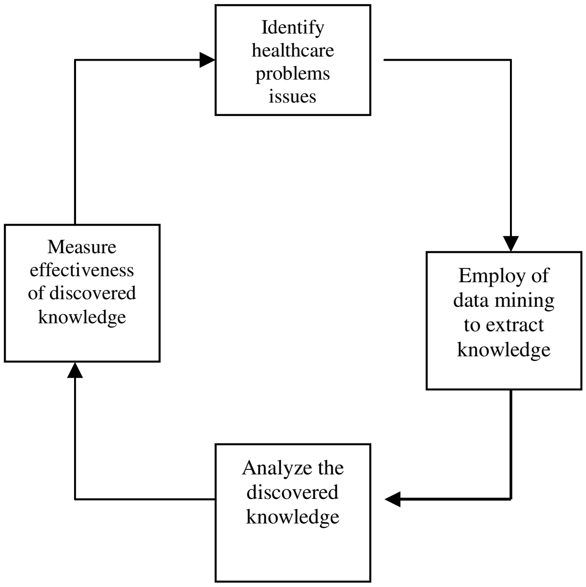
Identify healthcare problems issues
Employ of data mining to extract knowledge
Analyze the discovered knowledge
Measure effectiveness of discovered knowledge

Fig. 1: Data mining cycle

Massive healthcare data needs to be converted into information and knowledge, which can help control, cost and maintains high quality of patient care. Healthcare data includes Patient centric data and Aggregate data.

For health care organization to succeed they must have the ability to capture, store and analyze data (Fig. 1). Online analytical processing (OLAP) provides one way for data to be analyzed in a multi-dimensional capacity. With the adoption of data warehousing and data analysis/OLAP tools, an organization can make strides in leveraging data for better decision making[3].

Many healthcare organizations struggle with the utilization of data collected through an organization online transaction processing (OLTP) system that is not integrated for decision making and pattern analysis. For successful healthcare organization it is important to empower the management and staff with data warehousing based on critical thinking and knowledge management tools for strategic decision making. Data warehousing can be supported by decision support tools such as data mart, OLAP and data mining tools. A data mart is a subset of data warehouse. It focuses on selected subjects. Online analytical processing (OLAP) solution provides a multi-dimensional view of the data found in relational databases. With stored data in two-dimensional format OLAP makes it possible to analyze potentially large amount of data with very fast response times and provides the ability for users to go through the data and drill down or roll up through various dimensions as defined by the data structure.

With the widespread use of medical information systems including databases, there is an explosive growth in their sizes, Physicians are faced with a problem of making use of stored data. The traditional manual data analysis has become insufficient and methods for efficient computer assisted analysis indispensable. A Data Warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions[3]. A data warehouse is also often viewed as architecture constructed by integrating data from multiple heterogeneous sources to support structured and/or ad-hoc queries, analytical reporting and decision making.

Data mining techniques in health care: 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.

Classification data mining techniques: We now describe a few Classification data mining techniques with illustrations of their applications to healthcare.

Rule induction: is the process of extracting useful 'if-then' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents has the form

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. In the health care system it can be applied as follows:

(Symptoms) (Previous--- history) ------- > (Cause—of--- disease)

Rule Induction Method has the potential to use retrieved cases for predictions. The following example gives rule induction method for prediction blood alcohol concentration. The Table 1 shows ten attributes for alcohol measurement that were taken with a portable breadth testing machine (Alchosensor – III)[4].

Table 1: Attributes for alcohol measurement

Attributes	
Age (in yrs)	Meal (empty stomach, lunch, full)
Sex (M/F)	Amount of alcohol (ethanol in units)
Mass (in kg)	Blood_alcohol content (high/low)
Tobacco_use	Blood_pressure(high/low)
Height (in cm)	Time_duration (time spent drinking)

Using this technique, the attribute weight, sex, meal, time duration and amount produced the best results.

Example 1: If_then_rule induced in the diagnosis of level of alcohol in blood

IF Sex = MALE
AND Unit = 8.9
AND Meal = FULL
THEN
Diagnosis = blood_alcohol_content_HIGH.

Example 2: If_then_rule induced in the diagnosis of level of alcohol in blood

IF Unit > 9.1
AND Mass > 75
AND Meal = FULL
THEN
Diagnosis = blood_alcohol_content_HIGH

This method will predict whether the person is over the drink-driving limit based on the data taken from Table 1. We have applied this technique here because of the ready availability of subjects with some knowledge of the domain that can provide feedback on the explanations as shown in Example1 and 2. This can be used for decision making in healthcare.

Decision tree: It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree-based models which include classification and regression trees, are the common implementation of induction modeling[5]. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithm), CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, SPRINT[5].

The decision tree shown in Fig. 2 is built from the very small training set (Table 2). In this table each row corresponds to a patient record. 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.

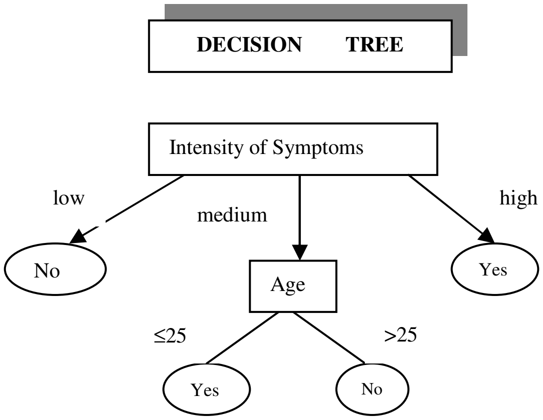

Fig. 2: A decision tree built from the data in Table 2

Table 2: Data set used to build decision tree of Fig. 2

Age	Gender	Intensity of symptoms	Disease (goal)
25	Male	medium	yes
32	Male	high	yes
24	Female	medium	yes
44	Female	high	yes
30	Female	low	no
21	Male	low	no
18	Female	low	no
34	Male	medium	no
55	Male	medium	no

Decision tree can be used to classify an unknown-class data instance with the help of the above data set given in the Table 2. The idea is to push the instance down the tree, following the branches whose attributes values match the instances attribute values, until the instance reaches a leaf node, whose class label is then assigned to the instance[5]. For example, The data instance to be classified is described by the tuple (Age=23, Gender=female, Intensity of symptoms = medium, Goal =?), where "?" denotes the unknown value of the goal instance. In this example, Gender attribute is irrelevant to a particular classification task. The tree tests the intensity of symptom value in the instance. If the answer is medium; the instance is pushed down through the corresponding branch and reaches the Age node. Then the tree tests the Age value in the instance. If the answer is 23, the instance is again pushed down through the corresponding branch. Now the instance reaches the leaf node, where it is classified as yes.

Artificial neural network (ANN): is a collection of neuron –like processing units with weight connections between the units. These models mimic the human brain and learn the patterns of a data set in order to make predictions.

Artificial Neural Networks (ANN) are analytical techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations from previous observations after executing a process called learning from existing data[6]. Neural networks or artificial neural networks are also called connectionist system, parallel distributed systems or adaptive systems because they are composed by a series of interconnected processing elements that operate in parallel as shown in Fig. 3. A neural network can be defined as 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. If the input to neuron is excitatory, it is more likely that this neuron connected to it. Neural networks are good for clustering, sequencing and predicting patterns but their drawback is that they do not explain how they have reached to a particular conclusion.

Artificial Neural Network is one of many data mining analytical tools that can be utilized to make predictions on key healthcare indicator such as cost or facility utilization. Neural networks are known to produce highly accurate results and in medical applications, can lead to appropriate decisions.

Artificial Neural networks are well suited to tackle problems that people are good at solving, like prediction and pattern recognition. Neural networks have been applied within the medical domain for clinical diagnosis, image analysis and interpretation[7,8], signal analysis and interpretation and drug development[9].

Artificial neural networks (ANN) 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, ANNs have been used to analyze blood and urine samples, track glucose levels in diabetics, determine ion levels in body fluids and detect pathological conditions[10]. A neural network has been successfully applied to various areas of medicine, such as diagnostic aides, medicine, biochemical analysis, image analysis and drug development[8]. Table 3 gives references of some of the medical applications of Neural Networks.

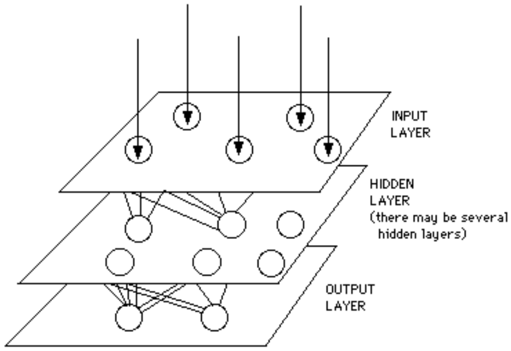

Fig. 3: A simple neural network diagram

Table 3: Applications of neural networks in various medical fields

Applications	References
Image Analysis	
Radiography (chest)	[11]
Radiography of heart	[12]
Image analysis of breast cancer nuclei	[13]
Image analysis of bladder carcinoma	[14]
Analysis of wave forms	
ECG	[15]
Clinical diagnosis	
Cervical cancer	[16]
Tumors	[17]
Retina damage classification	[18]
Analysis of side drug effects	[19]
Outcome Prediction	
Anesthesia	[20]
Breast cancer	[21]
Dental applications	[22]

Table 4: Step by step approach of classification to reduce risk factor

Steps	Decisions
S_1	Use the data (historical), the volume of data continuous to grow
S_2	Label the records of a patient with a particular disease (Class Attributes)
S_3	Develop a classification model from these records
S_4	The model so build may identify new factors which could detect the disease

Classification techniques in healthcare: The objective of the classification is to assign a class to find previously unseen records as accurately as possible.

If there is a collection of records (called as training set) and each record contains a set of attributes, then one of the attributes is class. The motive is to find a classification model for class attributes, where a test set is used to determine the accuracy of the model.

Table 5: Investigation of patient data

Physical Examination							Chemical/Microscopic Examination		
Code	Sex	Age	Disease	Colour	----------	Reaction (pH scale)	Glycated Haemoglobin (HbAlc level)	------- -	Microalbuminuria
1	M	12	5	P.Yellow		Acidic(5.5)	10.00		Yes
2	F	9	4	Yellow		Alkaline(8.5)	5.67		No
3	M	15	6	Golden		Acidic(6.0)	7.28		Yes
4	F	10	4	P.Yellow		Neutral (7.0)	8.00		Yes
----	----	---	----	-------		----------	------		----
----	----	----	----	-------		----------	-----		----
-----	----	----	----	-------		----------	------		----
100	M	16	4	Bright Red		Alkaline(7.45)	9.67		Yes
101	F	13	3	Straw		Alkaline(7.35)	9.00		Yes
102	M	1	2	Yellow		Neutral (7.0)	5.98		No
103	F	8	4	Yellow Green		Acidic(5.5)	4.39		No

The given data set is divided into training and test sets. The training set used to build the model and test set is used to validate it[5]. Classification process consists of training set that are analyzed by a classification algorithms and the classifier or learner model is represented in the form of classification rules. Test data are used in the classification rules to estimate the accuracy. The learner model is represented in the form of classification rules, decision trees or mathematical formulae. Table 4 illustrates a step by step approach of classification:

Decision trees can be used to classify new cases. They can construct explicit symbolic rules that generalize the training cases (rule induction and decision tree induction). New cases can then be classified by comparing them to the reference cases. Classification method can also be applied on Digital mammography images (for tumor detection in breast cancer)[23] to predict a class of categories (normal, benign or malign).

RESULTS AND DISCUSSION

Data mining of diabetes data: We present a case study of application of data mining and analyze data of children with Diabetes mellitus and Diabetes insipidus. The concept of Classification method has been applied in the study of Diabetes. Diabetes is a opportune disease for data mining technology for a number of factors, the huge amount of data is there and diabetes is a common disease that costs a great deal of money. Diabetes is a disease that can produce terrible complication such as thus blindness, kidney failure and premature cardiovascular death. Healthcare administers would like to know how to improve outcomes as much as possible.

In this we diagnosed about diabetes mellitus. There are two main types of diabetes mellitus. Type-1 (insulin-dependent) occurs before age 30, although it may strike at any age. The person with this type is usually thin and needs insulin injections to live and dietary modification to control his or her blood sugar levels. Type-2 (non-insulin dependent) occurs in obese adults over age 40. It is treated with diet and exercise, the blood sugar level is lowered with drugs[24,25].

Children with insulin-dependent diabetes mellitus of Type-1 were diagnosed. Type-1 (insulin dependent) diabetes mellitus is a chronic disease of the body metabolism characterized by an inability to produce enough insulin to process carbohydrates, fat and protein efficiently. Treatment of this disease requires insulin injection.

Table 6: Attributes of diabetes data

Attributes	Values
Number of times pregnant	-------
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	-------
Diastolic blood pressure (mm Hg)	-------
Triceps skin fold thickness (mm)	-------
2-Hour serum insulin (mu U/ml)	-------
Body mass index (weight in kg/(height in m)^2)	-------
Diabetes pedigree function	-------
Age (years)	-------

The above dataset as shown in the Table 6 is created by George John and appears on the UCL ML Data Repository at http//kddics.uci.edu. It contains 8 continuous attributes and 768 instance and two classes (one decision attribute) that determine either a person is or not having diabetes mellitus[26]. These attributes can filter out into Table 7 which is illustrated below. The Excerpt of Patient data with the results of Physical, Chemical and Laboratory examination is shown in Table 5. Final attributes and their values are presented in Table 7. Out of nine condition attributes, six attributes describe the result of physical examination, rest of the attributes of Chemical examinations. There are nine condition attributes in the laboratory examination report and one decision attribute i.e. micro-albuminuria[24].

The former six attributes include code, sex and age at which the disease was diagnosed. The decision attribute describe the presence of micro-albuminuria. The above results influence incidence of microalbuminuria in children suffering from diabetes type-I. All this information is stored in a database and is applied for the treatment of diabetes mellitus. Attributes available for each child were sex, age in years at diagnosis (<7, 7-12, 13-15, >15), disease duration in years (<6, 6-10, >10), previous history (yes/no), hypertension (yes/no), type of insulin used (reference range of blood glucose) (yes/no), Glycated Haemoglobin (HgbA1c) (<7, 7-10, >10) and micro-albuminuria (yes/no).

Table 7: Attributes and their values

Symbol	Attribute	Attribute value
I_1	Lab Ref. no.	Numeric value
I_2	Sex	M, F
I_3	Age of disease diagnosis(yrs)	< {range}>
I_4	Previous history	Yes, No
I_5	Duration of disease(yrs)	< {range}>
I_6	Diastolic Blood pressure (hypertension)	high, low
I_7	Mass	< {range}>
I_8	Sugar level	high, low
I_9	Reference range of Blood Glucose	high, low
X	Microalbuminuria	Yes, No

Given patient records with corresponding diagnosis, data mining methods are able to diagnose new cases. For instance, in the domain of early diagnosis of diabetic nephropathy disease, the patient record of laboratory examination comprises of condition attribute (including decision attributes).

ACKNOWLEDGEMENT

We are grateful to Dr H K Garewal, Aditi Medical Center, Vaishali, Gzb., India for providing an access to pathological medical data and for his involvement in this domain.

REFERENCES

1. Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.
2. Glymour, C., D. Madigan, D. Pregidon and P. Smyth, 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.
3. Shams, K. and M. Frashita, 2001. Data Warehousing Toward Knowledge Management. Topics in Health Information Management, 21: 3.
4. Jones, A.W., 1990. Physiological Aspects of Breath-Alcohol Measurements. Alcohol Drugs Driving, 6:1-25.
5. Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.
6. Lu, H., R. Setiono and H. Liu, 1996. Effective data mining using neural networks. IEEE Trans. On Knowledge and Data Engineering, 5: 8.
7. Miller, A., B. Blott and T. Hames, 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464.
8. Miller, A., 1993. The application of neural networks to imaging and signal processing in astronomy and medicine. Ph.D. Thesis, Faculty of Science, Department of Physics, University of Southampton.
9. Weinstein, J., K. Kohn and M. Grever *et al.*, 1992. Neural computing in cancer drug development: Predicting mechanism of action. Science, 258: 447-451.
10. Stanfford, G.C., P.E. Kelley, J.E.P. Syka, W.E. Reynolds and J.F. Todd, 1984. Recent improvements in and analytical applications of advanced ion-trap technology. Intl. J. Mass Spectrometry Ion Processes, 60: 85-98.
11. Robinson, P.J., 1997. Radiology's Achilles's heel: Error and variation in the interpretation of the Roentgen image. Radiol. Brit. J.
12. Itchhaporia, D., P.B. Snow, R.J. Almassy and W.J. Oetgen, 1996. Artificial neural networks: Current status in cardiovascular medicine.
13. Schnorrenberg, F., C.S. Pattichis, C.N. Schizas, K. Kyriacou and M. Vassiliou, 1996. Computer aidded classification of breast cancer nuclei.
14. Choi, H.K., T. Jarkrans, E. Bengtsson, J. Vasko, K. Wester, P.U. Malmsrom and C. Bausch, 1997. Image analysis based carcinoma. Comparison of object, texture and graph based methods and their reproducibility.
15. Simon, B.P. and C. Eswaran, 1997. An ECG classifier designed using modified decision based neural networks.
16. Romeo, M., F. Burden, M. Quinn, B. Wood and D. McNaughton, 1998. Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer.
17. Ball, G., S. Mian, F. Holding, Allibone Ro *et al.*, 2002. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomakers. Bioinformatics, 18: 395-404.
18. Aleynikov, S. and E. Micheli-Tzanakou, 1998. Classification of retinal damage by a neural network based system.
19. Domine, D., C. Guillon, J. Devillers, J. Lacroix and J.C. Dore, 1998. Non linear neural mapping analysis of the adverse effects of drugs.
20. Sharma, A. and R.J. Roy, 1997. Design of a recognition system to predict movement during anesthesia. IEEE Transactions.

21. Einstein, A.J., H.S. Wu, M. Sanchez and J. Gil, 1998. Fractal characterization of chromatin appearance for diagnosis in breast cytology.
22. Brickley, M.R., J.P. Stepher and R.A. Armstrong, 1998. Neural networks: A new technique for development of support systems in dentistry.
23. Zaiane, Osmar R, Antonie Maria-luiza and A. Coman, 2001. Application of data mining techniques for medical image classification. Second Intl. Workshop on Multimedia Data Mining. In Conjuction with ACM SIGKDD Conf. San Francisco, USA, Aug. 26.
24. Kelling, D.G. and J.A. Wentworth *et al*., 1997. Diabetes mellitus. Using a database to implement a systematic management program. NC. Med. J., 58: 368-371.
25. Kopelman, P.G. and A.J. Sanderson., 1996. Application of database systems in diabetes care. Med. Inform., (London), 21: 259-271.
26. http://www.comp.nus.edu.sg/dm2/p-download.html.