Medical Disease Classification in Data Mining

Asha Baby[#1], Shalu Krishna G[*2], Veena S Babu[#3]

[1]Assistant Professor, Department Of Computer Science, Vimal Jyothi Engineering College,Kannur
[2]Assistant Professor, Department Of Computer Science, College Of Engineering Aranmula, Pathanamthitta,
[3]Assistant Professor, Department Of Computer Science, Mount Zion Engineering College, Pathanamthitta

Abstract -*Health care domain is flooded with huge amount of data that holds sensitive information pertaining topatients and their medical conditions. Medical data minin can help obtain latent patterns or actionable knowledge. Dat mining techniques can discover such latent patterns or hiddenrelationships among the objects in the medical data sources.This will give know how to ascertain the progression of disease over a period of time. As medical data sources contain set o observations that are made from time to time with clinical parameters, considering temporal dimension of the data a fundamental parameter can give valuable insights related t temporal nature of diseases*

I. INTRODUCTION

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. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Regrettably all doctors do not possess expertise in every subspecialty and moreover there is a shortage of resource person's at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together.

Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. efficient and accurate implementation of automated system needs a comparative study of various techniques available. 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 . 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. Anticipating patients future behavior on the given history is one of the important applications of datamining techniques that can be used in health care management. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer based information and/or decision support systems. Health care data is massive. It includes patient centric data resource management data and transformed data. Health care organizations must have ability to analyzed 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.The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from qualitative visualization of clinical data by demanding a more quantitative assessment of information with the supporting of all clinical and imaging data. For instance it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. Likewise, physicians can also confirm their findings with the conformity of other physicians dealing with an identical case from all over the world.

Clinical decisions are often made based on doctors intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. In particulars, data mining may accomplish class description, association classification,clustering, prediction and time series analysis.

II. CLASSIFICATION AND PREDICTION

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a

better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions. For example,we can build a classification model to categorize bank loan. applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation. Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk resident data.

III.BACKGROUND STUDY

A. What Is Classification? What Prediction?

A bank loans officer needs analysis of her data in order to learn which loan applicants are safeand which are risky for the bank.A marketing manager at All Electronics needs data analysis to help guess whether a customer with a given profile will buy a new computer. A medical researcher wants to analyze breast cancer data in order to predict which one of three specific treatments a patient should receive. In each of these examples, the data analysis task is classification, where a model or classifier is constructed to predict categorical labels, such as safe or risky for the loan application data; yes or no for the marketing data; or treatment A, treatment B, or treatment C for the medical data. These categories can be represented by discrete values, where the ordering among values has no meaning. For example, the values 1, 2, and 3 may be used to represent treatments A, B, and C, where there is no ordering implied among this group of treatment regimes. Suppose that the marketing manager would like to predict how much a given customer will spend during a sale at All Electronics.

This data analysis task is an example of numeric prediction, where the model constructed predicts a continuous-valued function, or ordered value, as opposed to a categorical label. This model is a predictor. Regression analysis is a statistical methodology that is most often used for numeric prediction, hence the two terms are often used synonymously. We do not treat the two terms as synonyms, however, because several other methods can be used for numeric prediction. Classification and numeric prediction are the two major types of prediction problems. For simplicity, when there is no ambiguity, we will use the shortened term of prediction to refer to numeric prediction.

IV.HOW DOES CLASSIFICATION WORK

Data classification is a two-step process, as shown for the loan application data of Figure showes below. (The data are simplified for illustrative purposes. In reality, we may expect many more attributes to be considered.) In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or learning from a training set made up of database tuples and their associated class labels. A tuple, X, is represented by an n-dimensional attribute vector, X = (x1, x2,...., xn), depicting n measurements made on the tuple from n database attributes respectively, A1, A2,...., An.1 Each tuple, X, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute. The class label attribute is discrete-valued and unordered. It is categorical in that each value serves as a category or class. The individual tuples making up the training set are referred to as training tuples and are selected from the database under analysis. In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects.

Because the class label of each training tuple is provided, this step is also known as supervised learning (i.e., the learning of the classifier is supervised in that it is told to which class each training tuple belongs). It contrasts with unsupervised learning (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance. For example, if we did not have the loan decision data available for the training set, we could use clustering to try to determine groups of like tuples, which may correspond to risk groups within the loan application data.

This first step of the classification process can also be viewed as the learning of a mapping or function, y = f (X), that can predict the associated class label y of a given tuple X. In this view, we wish to learn a mapping or function that separates the data classes. Typically, this mapping is represented in the form of classification rules, decision trees, or mathematical formulae. The rules can be used to categorize future data tuples, as well as provide deeper insight into the database contents. They also provide a compressed representation of the data. In the second step the model is used for classification. First, the predictive accuracy of the classifiers estimated. If we were to use the training set to measure the accuracy of the classifier, this estimate would likely be optimistic, because the classifier tends to overfit the data (i.e. During learning it may incorporate some particular anomalies of the training data that are not present in the general data set overall). Therefore, a test set is used, made up of test tuples and their associated class labels. These tuples are randomly selected from the general data set. They are independent of the training tuples, meaning that they are not used to construct the classifier.

V. CLASSIFICATION BY DECISION TREE INDUCTION

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, C5.0 (a later version ID3 algorithm). The decision tree shown in Fig.3 is built from the very small training set (Table 1). 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. Decision tree can be used to classify an unknown class data instance with the help of the above data set given in the Figure 4. 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. 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 th instance. If the answer is 23, the instance is again pushed down through the corresponding branch. Now the instance reached the leaf node, where it is classified as yes.

VI. CLASSIFICATION BY CASE-BASED REASONING

Case-based reasoning (CBR) classifiers use a database of problem solutions to solve new problems. Unlike nearest neighbour classifiers, which store training tuples as points in Euclidean space, CBR stores the tuples or cases for problem solving as complex symbolic descriptions. Business applications of CBR include problem resolution for customer service help desks, where cases describe product-related diagnostic problems. CBR has also been applied to areas such as engineering and law, where cases are either technical designs or legal rulings, respectively.

Medical education is another area for CBR, where patient case histories and treatments are use to help diagnose and treat new patients When given a new case to classify, a case-based reasoned will first check if an identical training case exists. If one is found, then the accompanying solution to that case is returned. If no identical case is found, then the case-based reasoner will search for training cases having components that are similar to those of the new case. Conceptually, these training cases may be considered as neighbours of the new case. If cases are represented as graphs, this involves searching for subgraphs that are similar to subgraphs within the new case.

The case-based reasoner tries to combine the solutions of the neighbouring training cases in order to propose a solution for the new case. If incompatibilities arise with the individual solutions, then backtracking to search for other solutions may be necessary. The case-based reasoner may employ background knowledge and problem-solving strategies in order to propose a feasible combined solution. Challenges in case-based reasoning include finding a good similarity metric (e.g., for matching subgraphs) and suitable methods for combining solutions. Other challenges include the selection of salient features for indexing training cases and the development of efficient indexing techniques. A trade-off between accuracy and efficiency evolves as the number of stored cases becomes very large. As this number increases, the case-based reasoner becomes more intelligent. After a certain point, however, the efficiency of the system will suffer as the time required to search for and process relevant cases increases. As with nearest-neighbor classifiers, one solution is to edit the training database. Cases that are redundant or that have not proved useful may be discarded for the sake of improved performance. These decisions, however, are not clear-cut and their automation remains an active area of research.

VII. PROBLEM STATEMENT

The proposed system is identifying reliable information in the medical domain stand as building blocks for a healthcare system that is up-to date with the latest discoveries. By using the tools such as ML techniques with keyword symptom search. In this research, focus on diseases and treatment information, and the relation that exists between these two entities. The main goal of this research is to identify the disease name with the symptoms specified and get the Relation that exists between Disease-Treatment and classify the information into cure, prevent, side effect to the user. This system also provide most searched disease information

VIII. CONCLUSION

People care deeply about their health and want to be, now more than ever, in charge of their health and healthcare. Life is more hectic than has ever been, the medicine that is practiced today is an EBM in which medical expertise is not only based on years of practice but on the latest discoveries as well.Tools that can help us manage and better keep track of our health such as Google Health and MicrosoftHealthVaul are reasons and facts that make people more powerful when it comes to healthcare knowledge and management.

REFERENCES (SIZE 10 & BOLD)

[1] S. Novichkova, S. Egorov, and N. Daraselia, MedScan, A NaturalLanguage Processing Engine for MEDLINE Abstracts Bioinformatics, vol. 19, no. 13, pp. 1699-1706, 2003.

[2] [2] R. Bunescu, R. Mooney, Y. Weiss, B. Scho lkopf, and J.Platt,Subsequence Kernels for Relation Extraction Advances inNeural Information Processing Systems, vol. 18, pp. 171-178,2006.

[3] [3] M. Craven, Learning to Extract Relations from Medline Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.

[4] M. Craven, Learning to Extract Relations from Medline Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.

[5] S. Ray and M. Craven, Representing Sentence Structure in Hidden Markov Models for Information Extraction Proc. Intl Joint Conf. Artificial Intelligence (IJCAI 01), 2001.

[6] P. Srinivasan and T. Rindflesch, Exploring Text Mining from Medline Proc. Am. Medical Informatics Assoc. (AMIA) Symp., 2002.

[7] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression Nature Genetics, vol. 28, no. 1, pp. 21-28, 2001.

[8] B.J. Stapley and G. Benoit, Bibliometrics: Information Retrieval Visualization from Co-Occurrences of Gene Names in MEDLINE Abstracts Proc. Pacific Symp. Biocomputing, vol. 5, pp. 526- 537, 2000.

[9] R. Bunescu, R. Mooney, Y. Weiss, B. Scho lkopf, and J. Platt, Subsequence Kernels for Relation Extraction Advances in Neural Information

[10] Processing Systems, vol. 18, pp. 171-178, 2006. J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, Automatic Extraction of Protein Interations from Scientific Abstracts.