International Journal of Engineering Research and Development
e-ISSN : 2278-067X, p-ISSN : 2278-800X, www.ijerd.com
Volume 2, Issue 8 (August 2012), PP. 05-07

A Fuzzy Based Association Mining Approach for Medical Disease Prediction

Neeru Anand, Dr. Rajendar Singh Chhillar
Department of Computer Science & Applications, M.D. University, India

Abstract- ***The huge amount of textual data in distributed medical sources combined with the obstacles involved in creating and maintaining central repositories motivates the need for effective distributed information extraction and mining techniques. In this paper we present Decision Making Strategies by applying Fuzzy Logic to the patient's data through the Clinical Guidelines to make all probable decisions about the possibility of any of peculiar disease. Prior to applying fuzzy logic we first extract meaningful patterns of various diseases from the raw clinical guidelines which serve as a reservoir of a database of all diseases by applying text mining on these guidelines.***

Keywords- ***Data mining, health care, fuzzy rule set, clinical guidelines, medical diagnosis.***

I. INTRODUCTION

The influence of data mining on the quality of Health Care cannot be understated. All Health Care organizations retain detailed and comprehensive records of patient data. Trends and patterns identified in these records can positively impact the quality of Health Care. The huge amounts of patient data, makes identification of these trends an arduous task. However data mining applications, built for this purpose, can make this very simple and produce efficient results.

There have been several cases, where application of data mining techniques, have helped resolve a problem in the health industry. For instance, data mining on pneumonia patient records in a hospital, showed that patients who were administered medication immediately on arrival responded better than patients who were not administered medication on arrival. In order to arrive at this conclusion the data mining application, used several inputs, such as the tests and other information of the patients who showed better medication results. Various relations were drawn between the inputs. One of these was the relation between the results and the time taken to administer medication after arrival. It was found that, shorter the time, better the result.

There were several other key issues that were addressed at this time. The data mining tests proved that several tests, which were largely extraneous, were conducted on the patients. These led to a delay in the administration of medication and thereby affected the recovery of the patient. To overcome this, a standardized plan was created to treat pneumonia patients. The identification of these associations between inputs and finding the resultant best outcome was possible only because of data mining techniques.

A. Mining the Data

Health care organizations store huge amounts of data in the form of patient databases. Trends in these databases can be identified using data mining practices, which sort and model the data in order to arrive at a conclusion. The data mining applications present the data in the form of data marts. This allows end users to choose the specific sets of data, which they want to be analyzed. The data in these data marts can then be presented using a graphical user interface, arranging the data into columns and rows.

In the Health care industry, however, the lack of standard clinical vocabulary has hindered the process of data mining to a certain extent. For example a simple term such as ‘hypertension’ can be expressed in various ways in health care. This could lead to unnecessary problems, during the process of data mining. The increase in the use of standardized terms will reduce the percentage of errors in the data mining process.

Cleaning the data before it can be mined is also an important step in the data mining process. In many Health care organizations, the mode of preparing patient reports can lead to a good deal of confusion. For instance, in a certain hospital, a report was prepared, before and after a patient went in for an X-ray check. This could be construed as two different reports, when analyzing the data and produce erroneous results. Further in certain organizations, in order to reduce the number of reports, a patients’ record contains only the name of the attending physician and not the names of other physicians consulted or tests performed at a later stage, leading to erroneous predictions.

The data mining effort thus requires the wholehearted participation of all health care personal to produce comprehensive and correct reports, which can be mined. Further, the number of input variables for the data mining application has to be determined correctly. The number of inputs should not be so large, that it produces not be limited to such an extent, that they produce biased results. Co-operation between the physicians and analysts is also recommended, since some of the results might be more easily understood by the health care personal.

II. LITERATURE SURVEY

In year 2000, Shusaku Tsumoto performed a work," Problems with Mining Medical Data". Thus, it is highly expected that data mining methods will find interesting patterns from databases as reuse of stored data and be important for medical research and practice because human beings cannot deal with such a huge amount of data. In this paper, we focus on the characteristics of medical data and discuss how data miners deal with medical data. In year 2004, Y. Alp Aslandogan performed a work," Evidence Combination in Medical Data Mining". We combine the beliefs of three classifiers: k-Nearest Neighbor (kNN), Naïve Bayesian and Decision Tree. Dempster's rule of combination combines three beliefs to arrive at one final decision. Our experiments with k-fold cross validation show that the nature of the data set has a bigger impact on some classifiers than others and the classification based on combined belief shows better overall accuracy than any individual classifier. We compare the performance of Dempster's combination (with differentiation-based uncertainty assignment) with those of performance-based linear and majority vote combination models. We study the circumstances under which the evidence combination approach improves classification. In year 2006, Wong Kok Seng performed a work," Collaborative Support for Medical Data Mining in Telemedicine". This paper will discussed an idea on how to overcome above mentioned issues and proposed a solution that can be served as the platform for future medical data sharing in telemedicine. The successful development of the working prototype will greatly enhance the functionality of existing data sharing in the hospital. At the same time, the tools and algorithms designed in this idea will helps to solve some of the data mining challenges. In year 2008, Hai Wang performed a work," Medical Knowledge Acquisition through Data Mining". Data mining has been widely considered as an effective tool for knowledge discovery. This paper discusses the important role of medical experts for medical data mining, and presents a model of medical knowledge acquisition through data mining. In year 2010, Zhao Yongyi performed a work," Intelligent Data Mining for Economic Prediction and Analysis". This paper describes importance that the application of economic data in the data mining algorithm and its application, which combines with the current economic data of national macro-economic indicators, presents the data warehouse model structure and its implementation characteristics, and uses SQL Server 2005 data warehouse and data mining solutions on economic data for the application of data mining solution, system architecture, algorithms implementation, and finally discusses the application of data mining algorithms development trends and key technologies in the economic field. In year 2010, Shiguo wang performed a work," A Comprehensive Survey of Data Mining-based Accounting-Fraud Detection Research". Bayesian network, and stack variables etc. Regression Analysis is widely used on hiding data. Generally the detecting effect and accuracy of NN are superior to regression model. General conclusion is that model detecting is better than auditor detecting rate without assisting. There is a need to introduce other algorithms of no-tag data mining. Owing to the small size of fraud samples, some literature reached conclusion based on training samples and may overestimated the effect of model.

In year 2010, Mahnoosh Kholghi performed a work," Classification and Evaluation of Data Mining Techniques for Data Stream Requirements". Generally, two main challenges are to design fast mining methods for data streams and the need to promptly detect changing concepts and data distribution because of highly dynamic nature of data streams. The goal of this article is to analyze and classify the application of diverse data mining techniques in different challenges of data stream mining. For this goal, this article tries to categorize and analyze related researches for better understanding and to reach a framework that can map data mining techniques to data stream mining challenges and requirements.

III. PROPOSED SCHEME

Uncertainty plays a major role in the problem of guidelines representation. While natural languages (e.g., English) are quite suitable to express the uncertainty, present algorithmic languages call for precise recipes, and the translation from the first representation to the second presents a significant challenge. There are several types of uncertainty that may appear in clinical guidelines.

Relevance to a guideline may be available or has been collected, in which case an educated guess sometimes has to be made. Even if collected, the information can be unreliable. Second, it is non-specificity, connected with sizes of relevant sets. Frequently guidelines refer to other conditions, other risk factors, other significant conditions leaving it up to the doctor to decide what they are. To be translated into an algorithmic language, an explicit list of those conditions is required.

Third, it is the probabilistic nature of data and outcomes. There are few clinical signs that unequivocally point to a medical condition, and therefore to a predefined course of actions. Sensitivity and specificity of most clinical tests are far from ideal, and consequently they point to a likelihood, rather than presence or absence of medical condition. The outcome of any non-trivial recommendation is also, in a sense, a gamble. The words "usually", "likely", "commonly", "possibly", etc., express this type of uncertainty in natural languages.

Finally, it is fuzziness in determination of clinical signs that trigger the guidelines. It can be subjectivity in the assessment of a patient's symptoms, or in the interpretation of precise objective data, such as laboratory test results or even a patient's age. What exactly is the size of an "enlarged liver?" What exactly do we mean by "infants" or "middle-aged men?"

Fuzzy Set Theory (FST), introduced by Lotfi Zadeh in 1965, is the basis for Fuzzy Logic, Approximate Reasoning, Possibility Theory and other related disciplines. The main advantage of FST is that it allows transparency in knowledge representation. Formulation of decision rules mimics human thinking, and fuzzy logic permits one to construct fuzzy algorithms, flexible enough to represent the narratives of clinical guidelines. The key concept of FST is that of partial membership of elements in a set. In contrast to classical, "crisp" sets, where an element either belongs to the set or not, FST allows for degree of belonging to the set, usually real values taken from the range of 0 to 1, with 1 standing for complete membership and 0 for non-membership.

Fuzzy based association mining works on Boolean values which can be either true or false. For instance a patient suffering from high fever may be having temperature high then its truth value becomes 1 and if its false then its 0. Also if the value is intermediate such that it is neither true nor false then it takes the probability of both the condition. This whole approach we take into consideration for every attributes of a patient such that it becomes easy for the identification and the diagnosis of the disease. In such manner we classify the whole database using the fuzzy approach such as the age, weight, blood pressure and other medical terms using the probability of the truth and false values.

Medical diagnosis usually involves careful examination of a patient to check the presence and strength of some features relevant to a suspected disease in order to take a decision whether the patient suffers from that disease or not. A feature, like a runny nose for instance, may appear to be very strong for one patient but it can be moderate or even very light for another. It is the experience of the physician that tells him how to combine a set of symptoms (features and their strengths) to find out the correct diagnostic decision.

In the present work, we aim to capture the physician’s experience and store it in a set of tables. Inference is employed to develop a computer program that can automatically find out the certainty whether a patient having some specified symptoms suffers from any one of a set of suspected diseases. This certainty is specified as a crisp percentage value for every suspected disease.

We consider a set of *m* diseases ***D***, and define a collective set of *n* features ***F*** relevant to these diseases. Usually we have *n*>>*m*. Let:

$\boldsymbol{D} = \{d_{1,} d_{2,} d_{3,} \ldots, d_m\}$

$\boldsymbol{F} = \{f1_{,} f_{2,} f_{3,} \ \ldots , f_n\}$

To specify the symptoms of a patient, he would be checked against all features in the set ***F*** and a value would be assigned to each feature. The values are selected from the set:

{Very Low, Low, Moderate, High, Very High }

For example, a single symptom can be specified as < runny nose, Moderate >. By checking the patient for all *n* features of the set ***F*** and assigning a proper value for each feature, the set of patient’s symptoms ***S*** will be obtained as follows:

$\boldsymbol{S} = \{ <f_1 , v_1> , <f_2 , v_2> , <f_3 , v_3> , \ \ldots , <f_n , v_n> \}$

Where: v_i is the value assigned to the feature f_i when checking the patient, i=1, … ,n.

REFERENCES

[1]. Rafael S. Parpinelli, Heitor S. Lopes, Member, IEEE, and Alex A. Freitas.

[2]. Predictive data mining in clinical medicine: a focus on selected methods and Applications Riccardo Bellazzi, Fulvia, Ferrazzi and Lucia Sacchi.

[3]. Predictive data mining in clinical medicine: Current issues and guidelines by Riccardo Bellazzia,, Blaz Zupanb

[4]. DATA MINING FRAMEWORK by Hemambika Payyappillil, College of Engineering and Mineral Resources at West Virginia University.

[5]. FUZZY LOGIC IN CLINICAL DECISION SUPPORT SYSTEM by Jim Warren, Gleb Beliakov and Berend van der Zwaag.

[6]. Russell, S. and Norvig, P., Artificial Intelligence: A modern approach, pp. 23, Prentice-Hall International, 1995.

[7]. Buchanan, B. and Shortliffe, E., Rule-Based Expert Systems, Addison-Wesley, Reading, Massachusetts, 1984.

[8]. Waterman, D.A., A Guide to Expert Systems, Reading, MA: Addison-Wisley, 1986.

[9]. Kolodner, J. L., Case-Based Reasoning, California: Morgan Kaufman Publishers, 1993.

[10]. Phan, T. and G. Chen, Some Applications of Logic in Rule-Based Expert Systems, Expert Systems, vol.19, No.4, pp.208-223, 2002.

[11]. Clancey, W. J. and Shortliffe, E. H. (ed.)., NEOMYCIN: Reconfiguring a rule-based expert system for application to teaching. In: Readings in Medical Artificial Intelligence: The First Decade, Addison-Wesley, pp.361-381, 1984.

[12]. Zadeh, L. A., sets, Information and Control, 8, pp.338-353, 1995.

[13]. Leung R.W.K, Lau H.C.W., and Kwong C.K., On a responsive replenishment system: a logic approach, Expert Systems, vol. 20, pp. 20-32, 2003.