Prof. Mayura Kinikar, Harish Chawria, Pradeep Chauhan, Abhijeet Nashte / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.695-699

Data Mining in Clinical Practice Guidelines

Prof. Mayura Kinikar
Harish Chawria, Pradeep Chauhan, Abhijeet Nashte
Department Of Computer Engineering
Maharashtra Academy Of Engineering
Alandi, Pune

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. Recently, as the need to mine patterns across distributed databases has grown, Distributed Association Rule Mining (D-ARM) algorithms have been developed. These algorithms, however, assume that the databases are either horizontally or vertically distributed. In the special case of databases populated from information extracted from textual data, existing D-ARM algorithms cannot discover rules based on higher-order associations between items in distributed textual documents that are neither vertically nor horizontally distributed, but rather a hybrid of the two. In this paper we also present other 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*— Distributed Data Mining, Distributed Association Rule Mining, Clinical Guidelines, Guideline Interchange Format, Symptoms, Electronic Medical Record, Diagnosis, Fuzzy Logic.**

I. INTRODUCTION

Medical knowledge is vast and constantly changing, as well as expanding. The doubling time of medical knowledge is currently about 19 years, yet a recent survey found that textbooks available to physicians in their workplace were often more than 10 years old. Leaving aside both basic and specialized medical knowledge, a General Practitioner (GP) in Britain is expected to practice in accordance with the contents of numerous policies, referral protocols, government circulars, adverse drug effect warnings, etc. that form a stack. It is unrealistic to believe that the typical GP has read all these materials; it is a cognitive impossibility that all these rules are accurately analyzed on every patient when each consult lasts just several minutes.

As a result the whole data is Distributed throughout. There is a need to put on all these distributed results to integrate in a manner such that we can extract meaningful rules from this database. Based on these rules it becomes possible to diagnose a disease of a particular patient based on his personal information, his profile and the prolonged symptoms from which he is suffering from over a period or so. Thus our system works on a java platform which would finally output the probable disease to the patient by undergoing the above steps discussed and also the desired treatment which helps in diagnosing of that disease.

It has been recognized that simply making natural language clinical practice guidelines available on-line is not a complete solution to doctors' information management problems. The doctors must still know to seek out the right guideline information and take the time to find it. More significant practice improvements are achieved when guidelines are structured as algorithms that can trigger specific recommendations based on the content of an Electronic Medical Record (EMR). Comprehensive development of such algorithms, however, is frustrated because natural and relevant expressions of clinical guidance are apt to be somewhat imprecise in their context and phrasing.

Fuzzy Logic has a history of application for clinical problems including use in automated diagnosis, control systems, image processing and pattern recognition. Liu and Shiffman have demonstrated the application of fuzzy logic to model the imprecision of a published clinical practice guideline, which is cited by Zielstorff as a promising direction for future development of computer-based decision. This paper shows how fuzzy logic can fit the decision support framework and give a set of results to manage uncertainties.

Based on the dataset of the diseases or a domain we will perform detailed analysis by three algorithms. Such as Naive Bayes, K-means and Apriori algorithms.

II. CLINICAL PRACTICE GUIDELINES

Clinical practice guidelines (hereafter clinical guidelines or simply guidelines) are standardized specifications for care developed by a formal process that incorporates the best scientific evidence of effectiveness with opinions of experts in the fields. In general, they have been developed in an effort to reduce escalating health care costs without sacrificing quality and have been shown to improve health

care outcomes when followed. Many clinical practice guidelines are available for an extensive range of clinical problems, and several bodies have been established as clearing houses.

To be effective, guidelines need to be integrated into the physician's decision-making process in daily practice. The acceptance of guidelines by medical practitioners, however, depends on several factors, including awareness, availability, relevance, applicability in specific circumstances, mutual agreement, supporting evidence, etc.

Most current guidelines are implemented in printed form, or as direct translations of the printed text-based narratives, however there have been a number of attempts to provide effective electronic representations of clinical guidelines. In this process several key factors affecting their use have been established.

The highest probability of an effective guideline implementation occurs when patient-specific advice is provided at the time and place of a consultation. It has been recognized that the guideline statements should be linked to the actual patient data, and therefore be integrated into an electronic medical record. The most predictable impact is achieved when the guideline is made accessible through computer-based, patient-specific reminders that are integrated into the clinician's workflow. There are many obstacles on the way to making guidelines available in the form of patient-specific reminders. One such obstacle is the uncertainty and imprecision, inherent in clinical guidelines.

In general sense, regardless of its presentation, to be a collection of If…Then…rules, a diagram, a flowchart, a sequence of statements in a procedural language, etc. Usually clinical guidelines are implemented in the form of text narratives, describing possible medical conditions and signs with the appropriate recommendations. One profound reason we do not see guidelines represented as algorithms is that such narrative recommendations may not have traditional algorithmic representations. Some authors suggest that guidelines are not intended to be literally and directly applied, they specify a mixture of procedural and criterion-based knowledge, which the clinicians are tacitly expected to adjust and adapt according to the specific of a case. This fact creates a significant obstacle for computerizing clinical guidelines, their electronic exchange and assessment. Despite recent progress in developing formal syntax for guideline representation, in the computerised form the guidelines are mostly translations of text-based narratives.

III. UNCERTAINTIES IN GUIDELINES

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.

First, it is lack of information. Not every observation of 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.

IV. CLASSIFICATION BY NAIVE BAYES

Naive Bayesian classifiers have proven to be powerful tools for solving classification problems in a variety of domains. They have been successfully applied in the medical domain for solving diagnostic problems, such as the diagnosis of heart disease in newborn babies.

We use the dataset that in the attribute relationship format consisting of skin disease attributes and its relationship features. It is a typical dermatology database which shows four diseases such as psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, The dataset also has features related to this diseases.

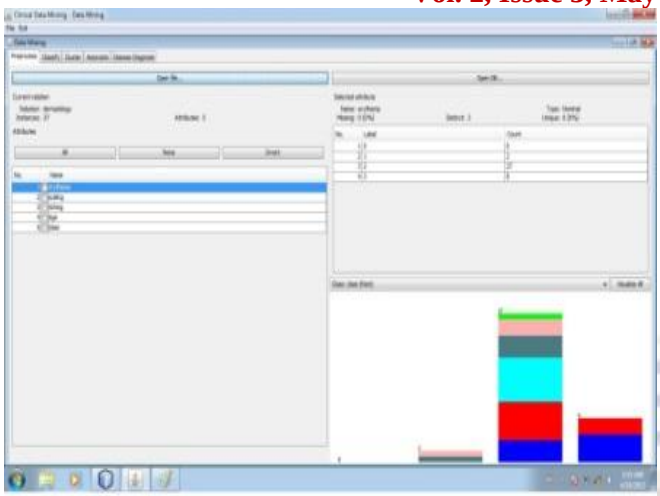

Figure-I

The above figure shows the snapshot of the execution of the project. The particular dataset which is stored on the hard disk is browsed and opened. On opening it shows the instances of the classes and the features as well as attributes as per the dataset of dermatology.

Constructing a naive Bayesian classifier starts with defining a class variable with its possible values and feature variables with their values.To complete the construction of a naive Bayesian classifier, various conditional probabilities have to be obtained. For each feature variable included in the classifier, more specifically, conditional probability distributions over its values given the different classes have to be defined. While the classifiers could be built from information provided in the literature, to establish their sensitivities, specificities and accuracies, a dataset was needed.

V. CLUSTERING BY K-MEANS

K-means is an algorithm to classify or to group your objects based on attributes or features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data. As we are using filter approach to get further and appropriate information related to the skin diseases from the dermatology dataset in order to get more precise result from the further apriori and fuzzy logic approach.

The basic step of k-means clustering is simple. In the beginning, we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids.

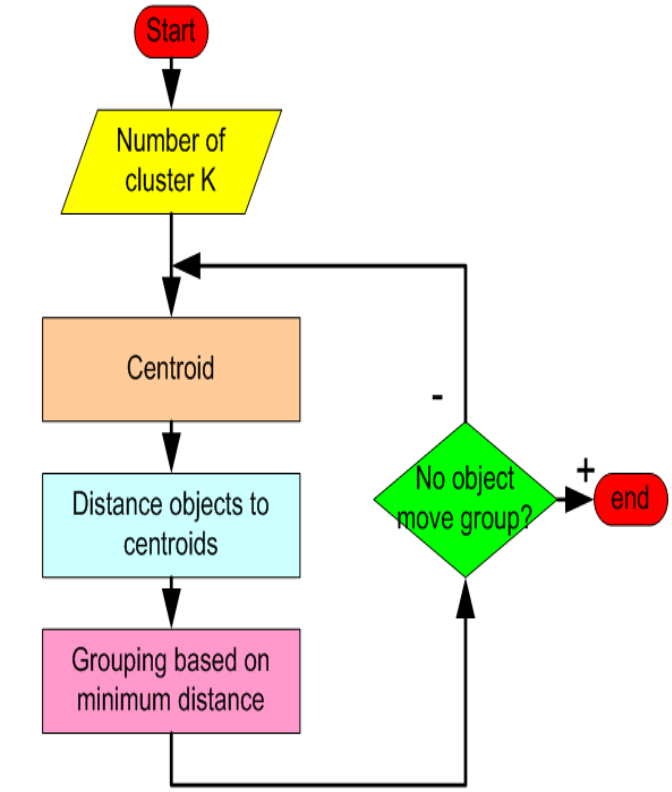

Figure-II

The above figure II shows the typical steps of how the clustering is done by k-means method. First, the total no of clusters is taken into consideration then the centroid of each cluster is calculated by constructing a distance matrix and further arithmetic operations are carried out. If the number of data is less than the number of cluster then we assign each data as the centroid of the cluster. Each centroid will have a cluster number. If the number of data is bigger than the number of cluster, for each data, we calculate the distance to all centroid and get the minimum distance. This data is said belong to the cluster that has minimum distance from this data.

Since we are not sure about the location of the centroid, we need to adjust the centroid location based on the current updated data. Then we assign all the data to this new centroid. This process is repeated until no data is moving to another cluster anymore. Mathematically this loop can be proved convergent. Thus the dataset is classifies further into clusters based on the attributes and features that are fed as input from the naive bayes classifier. K-means further filters the dataset and inputs it to diagnostic step.

VI. ASSOCIATION BY APRIORI ALGORITHM

Association rule mining (ARM) discovers associations between Items. Given two distinct sets of items, X and Y, we say Y is associated with X if the appearance of *X* implies the appearance of *Y* in the same context. ARM outputs a list of association rules of the format X Y, where X Y has a predetermined support and confidence. Many ARM algorithms are based on the well-known Apriori algorithm. In

Apriori, rules are generated from itemsets, which in turn are formed by grouping items that co-occur in instances of dataset.

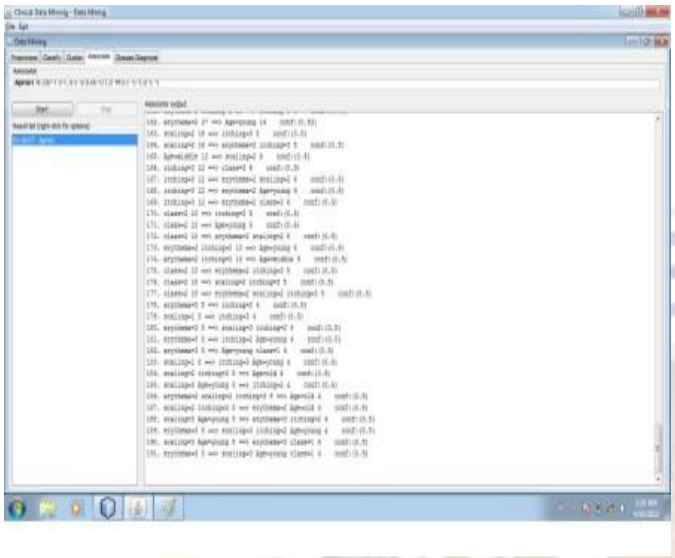

Figure III

The above figure III shows how association rules are formed on the classified dataset which is the output of the naïve bayes and k-means classifiers. These association rules are generated based on the features also known as symptoms such as erythema, scaling and itching of the four classes of skin diseases of the dermatology dataset.

The rules are discovered as per the instances occurred in the dataset. According to the discriminative features in the dataset the support and confidence of a particular disease is calculated with a predefined threshold values.

VII. DIAGNOSIS BY FUZZY LOGIC

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.

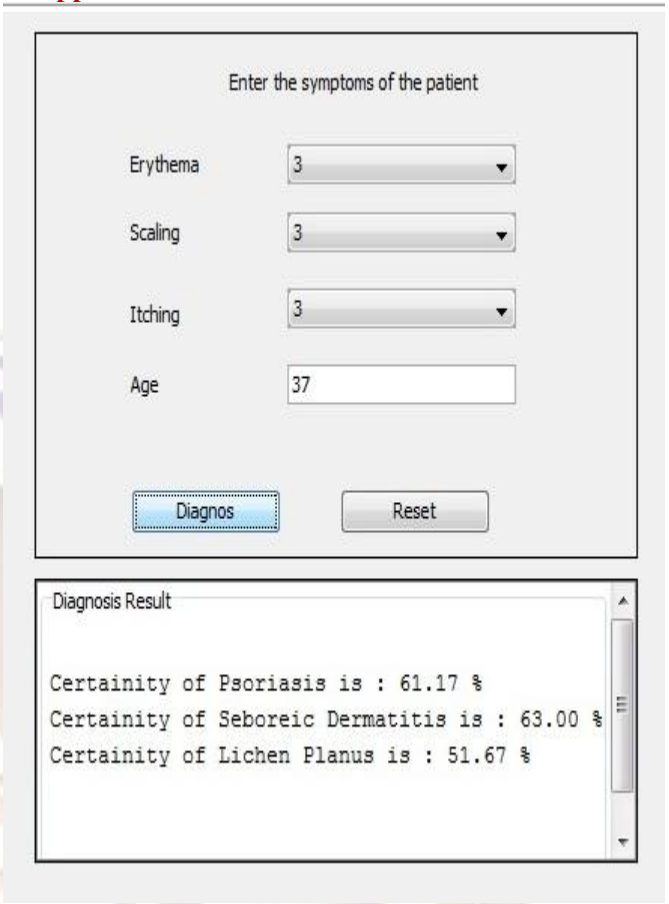

Figure IV

The above figure shows the final diagnosis step of the dataset of the dermatology which is fed as input. It goes through the various classification step as stated above in the paper. Finally by fuzzy logic which works on the Boolean values the decision is made on the rules discovered by association mining.

By stating the different prescribed values of the features of erythema, scaling and itching we can diagnose the certainty of the psoriasis and other skin infections. The probability is calculated in terms of percentage as per the instances, support and confidence of the classes of the features in the dataset.

VIII. CONCLUSION.

Clinical medicine is one of the most interesting areas in which data mining may have an important practical impact. The widespread availability of large clinical data collections enables thorough retrospective analysis, which may give healthcare institutions an unprecedented opportunity to better understand the nature and peculiarity of the undergoing clinical processes.

Combine the whole process of clinical process with computer generated treatment recommendations. Fast and Efficient implementation of clinical guideline reference for the medical practitioners.

.

IX. REFERNCES

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 issuesand guidelines by Riccardo Bellazzia,, Blaz Zupanb

4) DATA MINING FRAMEWORK byHemambika 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.