Study and Analysis of Predictive Data Mining Approaches for Clinical Dataset

Pooja Mittal
Department of Computer Science & Applications
Maharshi Dayanand University,
Rohtak 124001, Haryana, India

Nasib Singh Gill
Department of Computer Science & Applications
Maharshi Dayanand University,
Rohtak 124001, Haryana, India

ABSTRACT

Data Mining is an assortment of effective tool set to perform the statistical analysis on an immense dataset and to retrieve the valuable information from the dataset. In this work we have carried out an analytical survey on predictive data mining approaches on clinical dataset. The clinical dataset processing is one of the effective and most sensitive area which is studied under an expert environment. The present paper discusses KDD, data mining with reference to clinical expert system analysis, different applications and the approaches that can be used for the predictive data mining in same area. The scope of this paper is confined to the prediction of a person disease, based on symptoms dataset. The strength of data mining approaches in diverse clinical applications is also analyzed.

General Terms

Data Mining, KDD, Data Set, Disease, Prediction, Medical Diagnostic, Association Mining, Neural, Genetics, Fuzzy Logic, Regression Analysis, Markov Model, Decision Tree

Keywords

Clinical, Predictive, Expert System, Application, Mining Approaches

1. INTRODUCTION

Data Mining is itself a vital part of Knowledge Discovery in Databases (KDD) process, which infers knowledge from large set of databases coming from multiple sources. In today's demanding age, KDD is gaining popularity as a strong analytical solution for deriving useful information from voluminous data. KDD is an efficient process, as before extracting knowledge, it cleans the input data set to derive the desired reduced data set which produces the relevant knowledge to take constructive decisions.

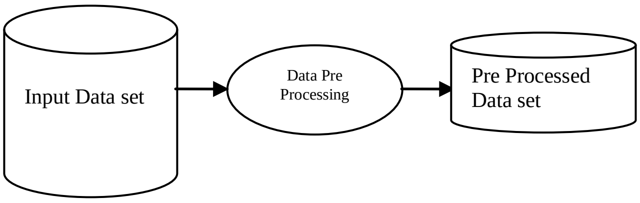

Fig 1: Data pre processing

As depicted in the Fig 1, large data sets undergo an immensely important process i.e. Data Pre-Processing, which intently extracts the useful required dataset which will be more condensed then input data set.

This process trim down the data set at two levels: First at attribute level & then on tuples. KDD provides number of methodologies for relevant attribute selection like Gain Ratio, Gini Index. For instance, consider a customer data set (approx. 5000 records) of any Pharmacist having following details: Bill_No, Name, Address, Medicine_Purch, Amount, Ph_No, Doctor_ Name, Email address and many others. In this data, fields like name, email, phone no can never be supportive in decision making process, thus by applying above stated methods, they are curtailed to reduce the number of attributes and the preprocessed table will have only the following fields: Bill_no, Address, Medicine_Purch, Amount, Doctor_Name.. Once attributes are picked then depending upon requirements, tuples are selected to further condense the data set. Once pre processing is done, Data mining process is applied to derive the desired knowledge as shown in Fig 2.

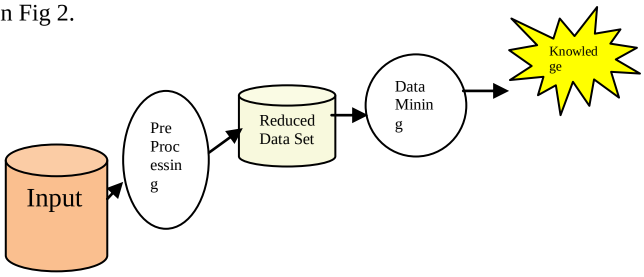

Fig 2: Knowledge Extraction Process

Data Mining is about to extract the hidden facts from a dataset and to discover the knowledge from it. This knowledge is then used to seize some short and long term decisions. These decisions are represented as the prediction. The mining used in such kind of estimation and analysis is called predictive data mining. The predictive mining is nothing just a collection of mathematical formulas to identify the hidden facts and trends. The predictive analysis removes the guess work and gives a scientific and analytical reason for the decision [1]. There are numerous applications of predictive analysis that includes the stock market prediction, weather forecasting, network attack analysis etc. In this present work we are discussing the predictive analysis respective to health care analysis. It is emphasized that in health care analysis, source and nature of data plays an incredible role. There are number of different approaches, tools and techniques that are supportive in prediction system. Some of these approaches are discussed in this paper.

Medical diagnostic is the foremost application that comes under the predictive data mining. But unlike other applications it is primarily concern with expert system. It means the study and analysis of medical diagnostic is possible under some medical expert of that field. Health care diagnostic prediction systems not only requisite the statistical and analytical study, it also requires the subjective study along with expert guidelines [2]. The accuracy level required in such system is very immense even though they cannot be implemented as an individual application without the expert concern. Such kind of system requires much more precise & accuracy in all steps of its architecture. The majority of the clinical database systems are dedicated, respective to the disease for which diagnostic is being performed. According to this the initial dataset is required. The acquired dataset must be secondary and taken from some physician or the concerned agency. The data collection in such system is again an expert system based application. The basic property of such dataset includes the heterogeneity of data set. Such kind of data is taken from patient lab tests, patient interviews, physician experience etc. All these approaches of data collection are used collectively to get a reliable and accurate dataset [3]. Once the dataset is acquired, it requires some filtration i.e. data cleaning. The cleaning is done under the expert of same disease. Once the cleaning is done the categorization of data is required. A rule based analysis is conceded to find the decision parameters on the basis of which the categorization will be performed. To achieve this, different clustering and classification approaches are performed. Once the categorization is done finally the decision making is performed in terms of prediction of disease.

2. HEALTH MONITORING APPLICATIONS

Though there can be numerous Health Monitoring Applications where data mining can play an incredible role for effective results, but most demanding applications are pictorially depicted in Fig 3 and discussed in the following section.

The most regular and the reliable application of data mining is health care system, which in turn keeps the historical information of patients, their symptoms, diagnostics, tests, lab results etc. While providing the diagnostic to a new patient the previous history is analyzed to verify or identify the diagnostic that can be applied to the patient. These kinds of systems are analytical system on which the statistical analysis is applied to acquire the summarized information from the dataset. Such kind of system is suggestion based system [4].

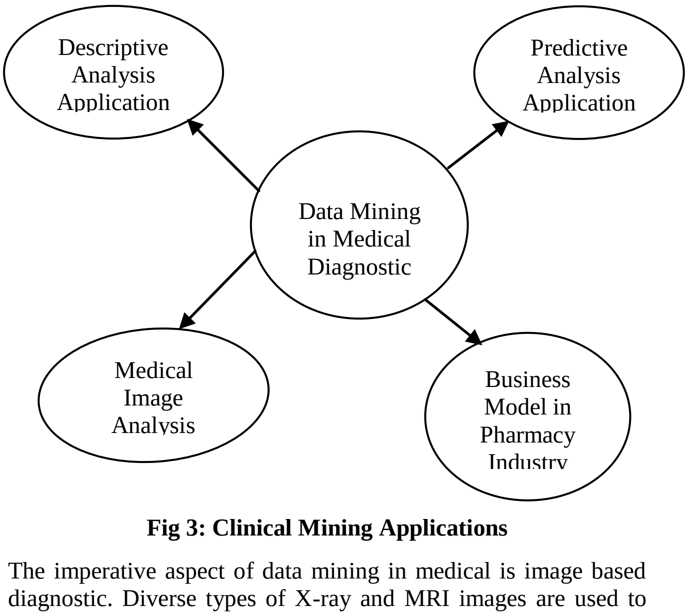
Descriptive Analysis Application
Predictive Analysis Application
Data Mining in Medical Diagnostic
Medical Image Analysis
Business Model in Pharmacy Industry

Fig 3: Clinical Mining Applications

The imperative aspect of data mining in medical is image based diagnostic. Diverse types of X-ray and MRI images are used to represent different kinds of tests, required for diagnosis. The mining approach is here worn to identify the kind of infection or the tumor in such medical images. In such systems, data mining improves the successful detection ratio as well as reduces the diagnosis cost of the patient. The successful detection of abnormality in such images is more then 90%. Even though such systems are implemented under the surveillance of some expert of same area [5].

The pharmacy industry is another prime area that employs data mining for designing the business strategies so as to decide the medicine that requires more productivity. This requires an analysis of demand and supply. The analysis also comprehends in estimating the assurance about the investment return. In such systems the medical diagnostic is analyzed in terms of convention of medicine relative to the particular disease. Different parameters are also considered to estimate the medicine usage by a particular age group, disease patients etc. It also includes the comparative study of other drugs on same disease and analysis of after effects [6]. In other words we can say in such application the medical diagnostic is mined as a business model.

3. APPROACHES FOR CLINICAL MINING

Data Mining is described as a statistical tool that can derive the analytical decisions on any well organized dataset. The reliability of the decisions driven depends on two major factors. First the data itself and other is the approach that is being implemented to extract the information. Here the data means how accurate and relevant the data is. The data itself is defined by several factors such as data size, medium from where data is fetched etc. Once the data is extracted, the next exertion is to filter it according to the requirements. Only the accurate and relevant data can conclude to good decisions. Once the required accurate data is available, the next job is the approach selection. The selection of approach depends on diverse factors such as linearity and impurities in dataset, size of data, operation being performed etc. In case of predictive analysis on clinical data the approaches that can be implemented are specified in this section.

3.1 Association Mining

Association Data Mining is one of the traditional data mining technique used for both the descriptive data mining and the predictive data mining. Association mining is basically the study of a decision parameter respective to the existence of other parameters [7]. In medical care application, the association is basically carried out between different parameters such as (i) between symptoms and disease (ii) between two different symptoms such as blood pressure and sugar level etc Association rules are generated as a result of association mining process, which in turn will perform the prediction. To ensure the quality of these generated rules, the foremost intended measures are support and the confidence.

Let's have two symptoms called S1 and S2 then the support rule S1 => S2 for a particular patient record is defined by support vector s. Here s is defined in percentage

$$\text{Support } (S1 \Rightarrow S2) = P\ (S1 \cup S2)$$

For the same rule the confidence level is defined as

$$\text{Confidence } (S1 \Rightarrow S2) = P\ (S2\ /\ S1)$$

3.2 Genetics

Genetics is the most accepted optimization algorithm that is being used by medical science to acquire the rapid and accurate results. A medical database is generally a vast dataset with number of decision parameters. In this genetic approach a smaller initial dataset is defined and each data value is represented as a chromosome. Now these chromosomes are processed by using genetic approach. The solution from these chromosomes is generated based on some rule set called fitness function. The chromosomes that follow this fitness function are processed further and rest values are discarded. By implementing the crossover operation on two different chromosomes new decision vectors are also generated. The GA is basically the search algorithm that uses the previous search as the base search and based on that, the result analysis is done in effective time [8]. Following query describes the complete GA process.

```
begin
  INITIALIZE population with random candidate solutions;
  EVALUATE each candidate;
  repeat
    SELECT parents;
    RECOMBINE pairs of parents;
    MUTATE the resulting children;
    EVALUATE children;
    SELECT individuals for the next generation
  until TERMINATION-CONDITION is satisfied
end
```

The approach followed by the genetic is shown in Fig 4.

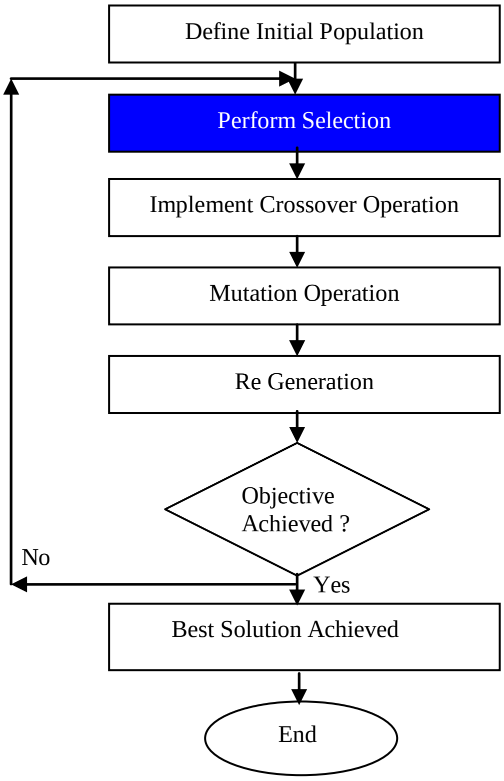

Fig 4: Genetics Process

3.3 Regression Analysis

Linear Regression is the basic type of regression analysis used in different statistical analysis applications. In this approach the analytical decision is taken about one variable or the parameter respective to other. The equational representation is given by

$$Y=b + w*X$$

Here b and w are the regression coefficient and X and Y are the related parameters. Changes in one parameter affect the other. Such as X can be the symptom and Y is the parameter related to disease. In simpler scenarios, linear regression may be useful but complex clinical analysis is not dependent on single vector it has number of correlated parameters that affect different output parameters. In such cases, multiple linear regressions is implemented [9].

3.4 Fuzzy Rule Based Decision Making

Some of the researchers used the fuzzy rule set as the decision making approach to design a clinical decision support system. Such kind of decision making is defined in terms of vectors called fuzzifier, fuzzy interference engine and defuzzifier. The first most significant task is the fuzzification in which the decision variables are considered that defines the input dataset along with membership function. Once the member ship functions are defined, the fuzzy rules are applied on it to perform the basic classification or the categorization. The fuzzy rule when associated with specific variable, defines the base rule which acts as the fuzzy interference engine [10]. On the basis of this base rule, Fuzzy interference engine performs the decision making. Finally to present the output as a quantitative value the defuzzification process is performed. In the clinical data set the member ship variables are in terms of some symptoms such as in case of heart disease the symptoms can be the heart beat, cholesterol level etc. The membership function is the specification of the range that decides the critical level of the value respective to the disease. The membership function is determined with the expert concern. Accordingly, the rule set will be generated and the decision making will be performed [11][12]. The fuzzy based interference system is a rule based classifier in which temporal algebra is used to generate the fuzzy rules. The fuzzy rule can be defined as a condition given as

$$F(\text{Condition}) \Rightarrow y$$

Here condition is the conjunction of attributes and y represents the class that supports the condition.

3.4 Neural Network Approach

Neural network is basically a classification technique used by copious decision making and the forecasting applications. In the medical field the neural network is extensively used to perform the predictive decision making under the defined set of rules [13]. The process performed by the neural is called the learning or the training of data. It means it derives the conclusion based results by deciding the decision vector variables and assigning specific weightage to these variables. It can process multiple variables collectively and effectively. Once the complete dataset is generated with weighted values it goes under the learning stage in which the weightage values are adjustified to predict the correct class. There are different kinds of neural network approaches comprising of the feed forward, back propagation, SVM, art network etc. Most of them perform the classification task. The major advantage of neural network is its parallel processing on multiple attributes and robustness in terms of weightage assignment. The neural network approach is the error prone approach that gives good results even in case of noisy data. In the clinical research area were the datasets are generally more clean and accurate the neural network is more effective and robust in terms of predicting the correct decision [14][15].

3.5 Markov Model

Hidden markov model is one of the pattern analysis tool used in many prediction based applications. Many of the recognition based applications uses the same approach to acquire the results with higher accuracy and the robustness. The major advantage of markov model is that it can work with both structured and unstructured attributes. In both cases the markov model identifies the patterns over the sequence of data [16]. Then these patterns are used to perform the decision making. To generate the pattern it uses distance analysis function with different vectors and discovers the value based patterns out of it. According to the distance vectors it defines various levels and with each level it reduces the dataset by eliminating the unsupported data values. Once it obtains the appropriate dataset it uses the association mining to conclude the final relationship [17]. In case of clinical dataset where we have outsized and accurate data values such model is more effective.

3.6 Decision Tree

Decision tree is one the major approach for knowledge representation. The decision trees are oriented to data classification and the prediction. The basic representation and construction of decision tree appears similar to the general tree structure in which data is organized in a parent child formation. Where each parent represents a class and the child nodes represents the data value that comes under that class. The decision tree algorithm exists where the hierarchy exists in class definition, respective to data values. The decision tree gives the benefit of hasty reduction of dataset for the processing [18]. Decision Tree algorithm provides the supervised data classification approach. Many of the classification Tools uses the same approach for data categorization [19].

As the Data Mining is a process of information as well as decision recovery that can be performed on large dataset. All above approaches are equally capable to discover the patterns from the large clinical datasets and to perform the prediction oriented decision making out of it. Some of these approaches are rule based approaches and some covers the concepts of soft computing techniques. Though all the approaches mentioned above are equally strong, but their applicability depends upon the nature and size of data as well as objective. Here association mining is the basic data mining approach where we have a precise set of data values. The approach can filter the dataset based on data values as well as the on the bases of association between attributes. The regression is also an analytical tool that check the relation between two or more attributes to perform the decision making. But as the number of attributes increases it is not much effective. The markov model is the prediction based analytical study that creates a group of values to perform the decision making. The decision tree is level based classifier that establishes the relationship between the classes itself. Because of this it reduce the data set very fast and return optimized processing to derive the fast results. But when the classification criteria overlap, this kind of approach cannot be used. Neural Network is another intelligent classifier performs the decision making based on the weight age assignment. Thus, every approach has its own arena and features.

4. CONCLUSION

In this paper potency of KDD & data mining is analyzed in terms of Health Monitoring Systems & study of data mining is performed on clinical dataset. Review of different predictive data mining approaches for the clinical dataset is presented in this paper. Some approaches are statistical that work on the user history and use basic data mining approaches to perform the disease prediction whereas other approaches are intelligent soft computing techniques that perform a classification based analytical study to identify the disease. Different statistical rule based and classification approaches are studied in the same vicinity. Strengths, process & weaknesses of every approach is studied and discussed. It is also analyzed that applicability of any approach is decided by the nature of data & application. This paper may act as a base map for selecting an appropriate data mining approach for diversified Clinical systems.

5. REFERENCES

[1] Debahuti Mishra, "Predictive Data Mining: Promising Future and Applications", Int. J. of Computer and Communication Technology, Vol. 2, No. 1, 2010

[2] E. Barati, M. Saraee, "A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction", Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI): March Edition, 2011

[3] Krzysztof J. Cio, "Uniqueness of medical data mining", Artificial Intelligence in Medicine 26 (2002) 1–24

[4] Yavar Naddaf, "Data Mining in Health Informatics"

[5] S. P. Deshpande, "Data Mining system and applications: A Review",InternationalJournal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010

[6] Jayanthi Ranjan, "Data mining in pharma sector:benefits", International Journal of Health Care Quality Assurance Vol. 22 No. 1, 2009 pp. 82-92

[7] Maragatham G, "a recent review on association rule mining", Maragatham G et al./ Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166 Vol. 2 No. 6 Dec 2011-Jan 2012

[8] Shital Shah, Andrew Kusiak, "Cancer gene searchwith data-mining and genetic algorithms", Computers in Biology and Medicine 37 (2007) 251 – 261

[9] Dave Smith, "Data Mining in the Clinical Research Environment", PhUSE 2007

[10] P.K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40

[11] Yo-Ping Huang, "Using Fuzzy Data Mining to Diagnose Patients' Degrees of Melancholia", Mobile Multimedia/Image Processing, Security, and Applications 2011

[12] U Keerthika, R Sethukkarasi, "A rough set based fuzzy inference system for mining temporal medical databases", International Journal on Soft Computing (IJSC) Vol.3, No.3, August 2012

[13] T.T.Nguyen, "Predicting CardioVascular Risk Using Neural Net Techniques"

[14] R. Sethukkarasi, "An Intelligent System for Mining Temporal Rules in Clinical Databases using Fuzzy Neural Networks", European Journal of Scientific Research ISSN 1450-216X Vol.70 No.3 (2012), pp. 386-395

[15] K. Usha Rani, "Analysis of heart diseases dataset using neural network approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011

[16] Ying Li, Sharon Lipsky Gorman, "Section Classification in Clinical Notes using Supervised Hidden Markov Model", IHI'10, November 11–12, 2010, Arlington, Virginia, USA

[17] Weiqiang Lin, "Temporal Data Mining Using Hidden

Markov-Local Polynomial Model",

[18] Fahad Shahbaz Khan, "Data Mining in Oral Medicine Using Decision Trees", International Journal of Biological and Life Sciences 2008

[19] D.Shanthi, Dr.G.Sahoo," Decision Tree Classifiers to Determine the Patient's Post-Operative Recovery Decision", International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (1): Issue (4)