Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

ID3 Algorithm Performance of Diagnosis For Common Disease

L.Sathish Kumar*
Department of Comp.sci & Engg,
Alagappa University,Karaikudi - 630001

Mrs.A.Padmapriya.M.C.A, M.Phil,(Ph.D)
Department of Comp.sci & Engg,
Alagappa University,Karaikudi - 630001
mailtopadhu@yahoo.co.in

Abstract- **Data mining is the process of collecting, searching through, and analyzing a large amount of data in a database, as to discover patterns or relationships. The diagnosis of disease ia a energetic and tricky job in madicine domain. The credit of common disease from dissimilar landscapes or marks is a multilayered problem that is not free false expectations and is recurrently attenreted by preciptate effects. Now day's people have been suffering from many common diseases like diabetes, pulmonary blockage, anemia, parasitic infectious diseases etc., at the same time people may get scared about their health conditions in a wide range. Hence they may take wrong decision regarding which disease they have been suffered from. This is kept in mind as a serious issue and we have used data mining as a tool to overcome this issue. Technically data mining can be used in finding co – relations at patterns in large types of fields, which is used in this paper for effective diagnosis of wide range at diseases along with their symptoms.**

Keywords: **ID3 Algorithm, Data mining, Common Disease , Pre-processing, Neural Network, Clustering, Decision Tree, Symptoms, Diagnosis.**

I .INTRODUCTION

Data mining technology provides a user oriented approach to novel and hidden patterns in the data. Day by day, rise lot of disease in the world, we cannot analysis all kind of diseases and how to take the correct medicine for all the diseases. This task is very difficult. The data mining techniques are very useful for finding the medicinal decision for the appropriate diseases. Data Mining has great potential for exploring the meaningful and hidden patterns in the data sets at the medical domain, these methods can be used for the medical and diseases diagnosis. Data Mining have two flavours directed and undirected. Directed data mining attempts to explain or categorize some particular target field such as income or response. Undirected data mining attempts to find patterns of similarities among groups of records without the use of a particular target field or collection of predefined classes. Data mining is largely concerned with building models. A model is simply an algorithm or set of levels that connects a collection of inputs to a particular target or outcome [1].Data mining makes the most sense when are large amounts of data. In fact, most data mining algorithms require large amounts of data in order to build and train the models that will then be used to perform classification, prediction, estimation, or other data mining tasks. Medical diagnosis is regarded as an important yet complicated task that needs to be executes accurately and efficiently. The automation at this system would be extremely advantageous. Regrettably all doctors do not posses expertise. In every subject specialist and more over there is a shortage and resource persons at certain places. Medical history data comprises of a number of tests

essential to diagnosis a particular disease[2]. Clinical databases are elements of the domain where the procedure of data mining has develop into an inevitable aspect due to the gradual incline of medical and clinical research data. It is possible for the healthcare industries to gain advantage of data mining by employing the same as an intelligent diagnostic tool. It is possible to acquire knowledge and information concerning a disease from the patient specific stored measurement as far as medical data is concerned. Therefore, data mining has developed into a vital domain in healthcare [3]. It is possible to predict the efficiency of medical treatments by building the data mining applications. Data mining can deliver an assessment of which courses of action prove effective [4] by comparing and evaluating causes, symptoms, and courses of treatments. The real-life data mining applications are attractive since they provide data miners with varied set of problems, time and again. Working on common disease patients databases is one kind of a real-life application. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process [5], [6].

This research work is the related of previous work [7] with intelligent and effective heart attack

© 2012, IJARCSSE All Rights Reserved

prediction system designed with the aid of neural network. In our previous work, we have presented an efficient approach for extracting patterns, which are significant to heart attack, from the heart disease data warehouses. In that we have utilized to common disease using data mining techniques: clustering and frequent pattern mining. Then clustering is performed on the pre-processed data warehouse using K-means clustering algorithm with K value so as to extract data relevant to common disease. Subsequently the frequent patterns significant to common disease diagnosis are mined from the extracted data using the ID3 algorithm.

Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all at 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. This paper aims to analyse the different predictive descriptive data mining techniques proposed in recent years for the diagnosis of common diseases.

The remaining sections of the paper are organized as follows: In Section 2, a brief review of some of the related works on common disease diagnosis is presented. An introduction about the common disease and its effects are given in Section 3. The extraction of significant patterns from common disease data warehouse is detailed in Section 4. The common disease prediction system designed with the aid of ID3 is elucidated in section 5. The experimental results are described in Section 6. The conclusions are summed up in Section 7.

II. REVIEW OF RELATED BACKGROUND LITERATURE

Due to the resource constraints and the nature of the paper, the main method used in this paper is the ID3 algorithm, which is used for the classification and the pattern reorganization technique. The paper is based on many of the survey of the journals and the publications in the fields of the data mining and in the medicinal field. This paper mainly focused on the classification and the pattern recognizes which is used to divide the user defined categories. Numerous works in literature related with disease diagnosis using data mining techniques have motivated our work. Some of the work discus following.

Predictive Data Mining For Medical Diagnosis [8]. This Paper Survey Of The Heart Disease Using Id3 Algorithms Of Naïve Bayes, Decision Tree, Nearest Neighbors Algorithms (K-Nn). This Survey A Gives The Report The Decision Tree Is The Highly Predictive For Heart Disease.

Decision Support System For Medical Diagnosis Using Data mining [9]. These papers referred by the survey of compare 3 disease are heart disease, diabetes, hepatitis. So this survey using algorithm was ID3, C4.5, and CART algorithms.

III COMMON DISEASES

The state of complete physical, mental and social well-being is called health. Disease is a disorder or malfunction of the mind or body, which leads to a departure from good health. Here following pages discuss some common disease .

Physical disease

Results from permanent or temporary damage to the body.

Malaria

Malaria infects an estimated 300 million people, and is spread by mosquitoes, transfusions, and shared hypodermic needles. Control of mosquito populations has led to declines in malaria in many areas.

Mental diseases

A disease that affects a person's mind Thoughts, emotions, memory and personal and social behaviour. May have physical symptoms.

Cholera

Cholera is an acute, diarrheal illness caused by infection of the intestine with the bacterium *Vibrio cholera*. The infection is often mild or without symptoms, but sometimes it can be severe. Approximately one in 20 infected persons has severe disease characterized by profuse watery diarrhoea, vomiting, and leg cramps. In these persons, rapid loss of body fluids leads to dehydration and shock.

Typhoid

Typhoid fever is contracted when people eat food or drink water that has been infected with Salmonella typhi. It is recognized by the sudden onset of sustained fever, severe headache, nausea and severe loss of appetite. It is sometimes accompanied by hoarse cough and constipation or diarrhoea. Case-fatality rates of 10% can be reduced to less than 1% with appropriate antibiotic therapy. Paratyphoid fever shows similar symptoms, but tends to be milder and the case-fatality rate is much lower.

IV. RESEARCH CONTRIBUTION

The patterns from the common disease data warehouse are presented in this section. The common disease data sets contains the screening clinical data of disease afire patient. Initially, the data set is pre-processed to make the mining process more efficient. The pre-processed data set is then clustered using the K-means clustering algorithm with K=2. This result in two clusters, one contains the data that are most relevant to common disease and the other contains the remaining data. The frequent patterns are mined from the data, relevant to common disease, using the ID3 algorithm. this paper proposed for effective diagnosis of wide range at diseases along with their symptoms.

© 2012, IJARCSSE All Rights Reserved

A .Data Pre-processing

Cleaning and filtering of the data might be necessarily carried out with respect to the data and data mining algorithms employed so as to avoid the creation of deceptive or inappropriate rules or patterns [10]. The actions comprised in the pre-processing of a data set are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. In order for making the data appropriate for the mining process it needs to be transformed. The raw data is changed into data sets with a few appropriate characteristics. Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm [11]. In our approach, the common disease data set is refined by removing duplicate records and supplying missing values. Furthermore it is also transformed to a form appropriate for clustering.

B. Clustering Algorithms

The categorization of objects into various groups or the partitioning of data set into subsets so that the data in each of the subset share a general feature, frequently the proximity with regard to some defined distance measure [12], is known as Clustering. The clustering problem has been addressed in numerous contexts besides being proven beneficial in many applications. Clustering medical data into small yet meaningful clusters can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques [13]. Numerous methods are available in the literature for clustering. We have employed the renowned K-Means clustering algorithm in our approach.

C. Overview of ID3 Algorithm

Itemized Dichotomozer 3 algorithm or better known as ID3 algorithm [14] was first introduced by JR.Quinlan in the late 1970's. It is greedy algorithm that selects the next attributes. The information gain associated with the attributes. The information gain is measured by entropy ID3 algorithm. Refers that the generated tree is shorter and the attributes with lower entropies are near the top of the tree.

ID3 Algorithm

Function ID3 (I, 0, T)
{
/* I is the set of input attributes
/* 0 is the output attributes
/* T is a set of training data
/* function ID3 returns a decision tree
*/ if (T is empty)
{
Return a single node with the value of the most frequent value 0 in T;
/* now handle the case where we can't return a single node compute the information gain for each attribute in I relative to T. Let x be attributes with largest gain (X, T) of the attributes in I;
Let {x, j/j=1, 2... m} be the values of x;
Let {t, j/j =1, 2... m} be the subsets of T;
When T is partitioned. According the value of x; Return a tree with root node labelled x an arcs labelled x_1,x_2,....., X _ m , where the 1 arcs go to the trees.
ID3(I-{X}; 0(T-1), ID3(I-{X},O , T2),..... ID3 {I-{X},O,T-M); }

V. COMMON DISEASE DIAGNOSIS SYSTEM USING ID3 ALGORITHMS

The reason of the choosing ID3 algorithm easily understanding table prediction rules is created from the training data. Builds the fastest tree. Build a short tree, only need to test enough attributes until all data is classification finding leaf nodes enables text to be pruned, reducing number of tests, whole dataset is searched to create tree. Mathematical algorithm for building the decision tree. Build the tree from the top down; with the back tracking information gain is used to select the most attributes for classification.

A. Common Disease Procedures

1.Select the dataset for which the test to be retrieved.
2.By using the ID3 algorithm sort the specific pattern and classify the datasets based on the symptoms.
3.Then pre-process the fields of dataset based on "Symptom" field and then diagnosis the causes and treatment of the disease also.
4. The paper focuses, the retrieval of dataset, based on the ID3 algorithm that result in the specific dataset fields retrieval.

B. Result Obtained from ID3 and Neural Network

Common Disease diagnosis System using data mining techniques, namely, ID3 ,Neural Network. is implemented in [15] using .NET platform .its Web-based, user-friendly, scalable, reliable and expandable system. It can also answer complex "what if" queries which traditional decision support systems cannot. Using medical profiles such as age, sex, symptoms, disease it can predict the likelihood of patients getting a Common disease diagnosis. It enables significant knowledge, e.g. patterns, relationships between medical factors related to common disease. As a Data source a total of 600 records with 6 medical attributes (factors) were obtained from the Cleveland common Disease database. Figure 1 lists the attributes. The records were split equally into two datasets: training dataset (300 records) and testing dataset (300 records) Table 2 summarizes the results of all six sample models. ID3 to be most effective as it has the highest percentage

© 2012, IJARCSSE All Rights Reserved

of correct predictions (94%) for patients with common disease,

followed by Neural Network (with a difference of less than 20%) , however, appears to be most effective for predicting patients with no common disease (74%) compared to the other one models.

Figure 1 : Accuracy of ID3 And Neural Network

Techniques	Accuracy
ID3 Algorithm	74%
Neural Network	46%

Figure 2 : Totally Accuracy Between ID3 and NN

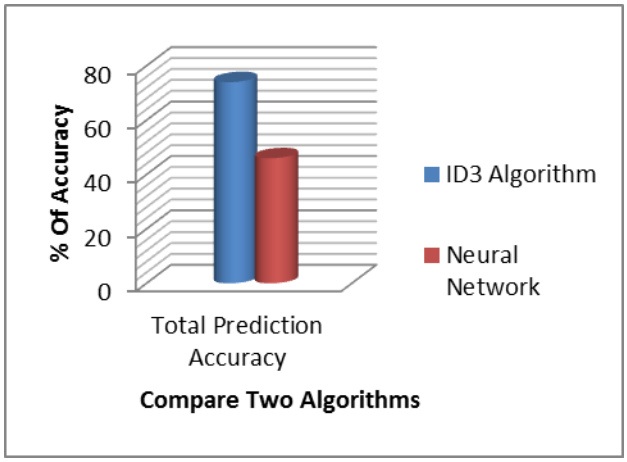
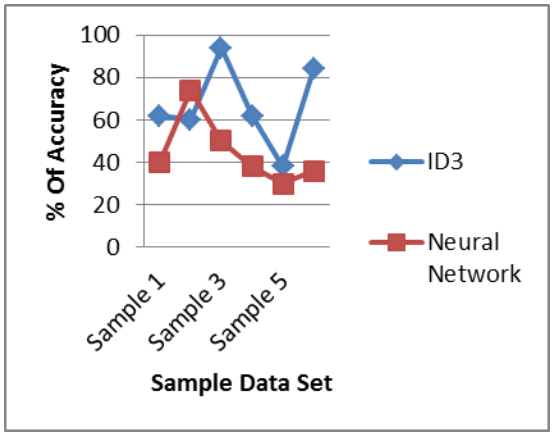


Figure 3 : Accuracy Chart Of Samples



VI. EXPERIMENTAL RESULT

This section presents the experimental results with the various classifications of the diseases based on the attribute classification. The results focus on the symptoms of the diseases and then, the classification of the diseases. The following figure shows the results of the classification and the association.

At the fig-4 the classification of the diseases is taken place. In the fig-5 the basic symptoms of the particular disease is defined. At the fig-6 the symptoms that are taken as input by the user is associated according to the attributes off the record. In the fig-7 the possibilty of disease that may cause according to the symptoms is distinct. Finally the selective treatment of the case is given by the classification and the association rule of the data mining.

The results show that by using the ID3 algorithm the classification of the attributes becomes easier hance varification also becomes easier. Therefore this rule cn be applied for mining and prediction.

Figure 4 : Diagnosis Step1

Figure 5 :*Diagnosis* Step2

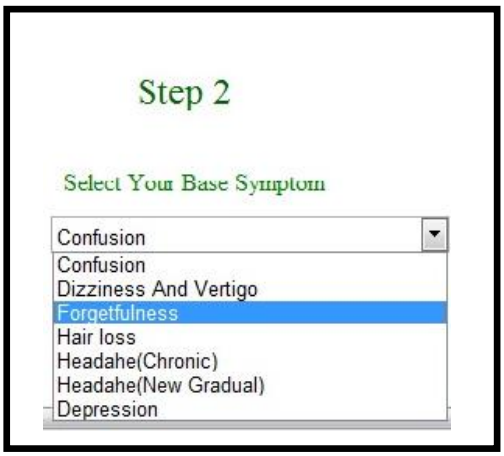


Figure 6: Diagnosis Step3

© 2012, IJARCSSE All Rights Reserved

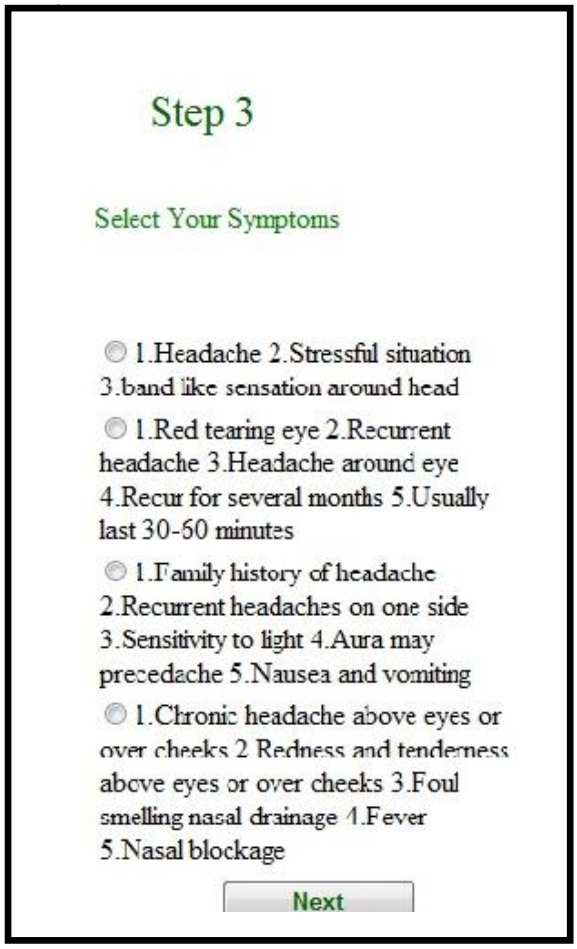
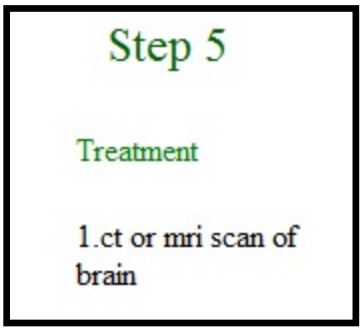


Figure 7: Diagnosis Step 4

Figure 8: Diagnosis Step 5



VII. CONCLUSION

In this paper the problem of constraining and summarizing two algorithms of data mining used in the field of medical prediction are discussed. The focus is on using two algorithms and combinations of several target attributes for intelligent and effective common disease diagnosis using data mining.

For diagnosis common disease, significantly 12 attributes are listed and with basic data mining technique other approaches e.g. Clustering and Association Rules, soft computing approaches etc. can also be incorporated. The outcome of diagnosis data mining technique on the same dataset reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of ID3 but other predictive methods like Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the ID3 and Neural Network.

In this paper we proposed the procedure for retrieval of dataset with, relevant fields using ID3 algorithm. Unlike previous works it is based on the individual diagnosis for specific symptoms of the disease. This paper concluded with the individual retrieval of dataset that predicts the diagnosis on the whole.

VII. REFERENCE

[1] Sally Jo Cunningham and Geoffrey Holmes, "Developing innovative applications in agriculture using data mining", In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, 1999.

[2] Anamika Gupta, Naveen Kumar, and Vasudha Bhatnagar, "Analysis of Medical Data using Data Mining and Formal Concept Analysis", Proceedings Of World Academy Of Science, Engineering And Technology,Vol. 6, June 2005,.

[3] S Stilou, P D Bamidis, N Maglaveras, C Pappas, “Mining association rules from clinical databases: an intelligent diagnostic process in healthcare”,

[4] Hian Chye Koh and Gerald Tan, "Data

© 2012, IJARCSSE All Rights Reserved

Mining Applications in Healthcare", Journal of healthcare information management, Vol. 19, No. 2, pp. 64-72, 2005.

[5] Frank Lemke and Johann-Adolf Mueller, "Medical data analysis using self-organizing data mining technologies," Systems Analysis Modelling Simulation, Vol. 43, No. 10, pp: 1399 -1408, 2003

[6] Andreeva P., M. Dimitrova and A. Gegov, "Information Representation in Cardiological Knowledge Based System", SAER'06, pp: 23-25 Sept, 2006.

[7] Shantakumar B. Patil, Y.S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", IJCSNS International Journal of Computer Science and Network Security, Vol. 9 No. 2 pp. 228-235, February 2009.

[8]. Sunitha soni, Associate professor PREDICTIVE DATAMINING FOR MEDICAL DIAGNOSIS AN OVERVIEW OF HEART DISEASE March 2011.

[9] D.Senthil Kumar, DECISION SUPPORT SYSTEM FOR MEDICAL DIAGNOSIS USING DATA MINING march 2011.

[10] Gerhard Münz, Sa Li, and Georg Carle, "Traffic anomaly detection using k-means clustering", In Proc. Of Leistungs-,Zuverlässigkeits-und Verlässlichkeitsbewertung on Kommunikationsnetzen und Verteilten Systemen, 4. GI/ITG-Workshop MMBnet 2007, Hamburg, Germany, September 2007.

[11] Wynne Hsu, Mong-Li Lee, Bing Liu, Tok Wang Ling, "Exploration mining in diabetic patients databases: findings and conclusions", KDD 2000: pp: 430-436, 2000

[12] Zakaria Nouir, Berna Sayrac, Benoît Fourestié, Walid Tabbara, and Françoise Brouaye, "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering," Journal of Communications, Vol. 2, No. 6, pp. 30-37, November 2007.

[13] F. H. Saad, B. de la Iglesia, and G. D. Bell, "A Comparison of Two Document Clustering Approaches for Clustering Medical Documents", Proceedings of the 2006 International Conference on Data Mining (DMIN-06), 2006.

[14]. Quinlan J.R., INDRODUCTION OF DECISION TREES Machine learning. VOL 1986, 81-106

[15] Sellappan Palaniappan Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008

[16] Sunita Soni, O.P.Vyas, Using Associative Classifiers for Predictive Analysis in Health Care Data Mining,International Journal of Computer Application (IJCA, 0975 –8887) Volume 4– No.5, July 2010, pages 33-34.

[17] .Sunitha soni, Associate professor PREDICTIVE DATAMINING FOR MEDICAL DIAGNOSIS AN OVERVIEW OF HEART DISEASE March 2011.

[18]., D.Senthil Kumar, DECISION SUPPORT SYSTEM FOR MEDICAL DIAGNOSIS USING DATA MINING march 2011.

.[19] .Asha Rajkumar, G.Sophia Reena, Diagnosis Of Heart Disease Using Data mining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver.1.0 September 2010.

© 2012, IJARCSSE All Rights Reserved