An Efficient Algorithm for Disease Prediction with Multi Dimensional Data

Smitha.T
PhD- Research Scholar
Karpagam University,
Coimbatore
(Asst.Professor-SNGIST, N.Paravoor)

V.Sundaram, PhD.
Director-MCA
KEC
Coimbatore.

ABSTRACT

The main objective of this study is to create a fast, easy and an efficient algorithm for disease prediction , with less error rate and can apply with even large data sets and show reasonable patterns with dependent variables.
For disease identification and prediction in data mining a new hybrid algorithm was constructed. The Disease Identification and Prediction(DIP) algorithm, which is the combination of decision tree and association rule is used to predict the chances of some of the disease hits in some particular areas. It also shows the relationship between different parameters for the prediction. To implement this algorithm using vb.net software was also developed.

Keywords: association rule, clustering, decision tree, multidimensional data

1. INTRODUCTION

Data mining in databases is the automatic extraction of implicit and interesting patterns from large data collections. Data mining is a field at the intersection of computer science and statistics and is the process that attempts to discover patterns in large data sets.. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use[1].
Data mining is a system of searching through large amounts of data for patterns. The main goal of data mining is to extract important information from data that was not previously known. It is commonly used to recognize certain patterns or trends. One important factor of data mining is that it will often be used to analyze information from a variety of different perspectives. The important information that is gained from data mining can be used to increase profits or lower costs. Data mining is a logical process that is used to search through large amounts of information in order to find important data. The goal of this technique is to find patterns that were previously unknown. Once we have found these patterns, we can use them to solve a number of problems.[9]
The goal of the person who uses data mining is , he/she should be able to predict certain behaviors or patterns. Once the user is able to predict the behavior of something which he analyzing, he will be able to make strategic decisions that can allow him to achieve certain goals.

2. LITERATURE REVIEW

Behrouz Minaei-Bidgoli, Elham [1] explained a new approach of using data mining tools for customer complaint management using association rule mining . The data of citizens' complaints on Tehran municipality were analyzed. Using this technique it was possible to find the primary factors those caused complaints in different geographical regions in different seasons of the year. The idea of contrast association rules were also applied to discover the variables that influence complaints occurrence. In order to accomplish this objective, citizens were grouped according to the demographical and cultural characteristics and the contrast association rules were extracted. The results show that there is a strong relationship between citizen gender and education and patterns of complaints occurrence.

K.Srinivas et al. [2] ,in their study, briefly examined the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data enables significant relationships between medical factors related to heart disease. In this paper, the authors have presented an intelligent and effective heart attack prediction methods using data mining. Firstly, they have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack Based on the calculated significant weight age, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Five mining goals are defined based on business intelligence and data exploration. The goals are to be evaluated against the trained models. All these models could answer complex queries in predicting heart attack.

Sunita Soni, Jyoti Soni, Ujma Ansari [3] analised the problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction are discussed. The focus is on using different algorithms and combinations of several target attributes for intelligent and effective heart attack prediction using data mining. For predicting heart attack, significantly fifteen attributes are listed and with basic data mining technique other approaches like Time Series, Clustering and Association Rules, soft computing approaches etc. can also be incorporated. The outcome of predictive data mining technique on the same dataset reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.
Shelly Gupta et al. [4] summarized various review and technical articles on breast cancer diagnosis and prognosis. In this paper the authors, present an overview of the current research carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis. From the above study it is

observed that the accuracy for the diagnosis analysis of various applied data mining classification techniques are highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. The prognostic problem is mainly analyzed under ANNs and its accuracy came higher in comparison to other classification techniques applied for the same. But more efficient models can also be provided for prognosis problem like by inheriting the best features of defined models. In both cases we can say that the best model can be obtained after building several different types of models, or by trying different technologies and algorithms.

G.V. Nadiammai, S.Krishnaveni Dr.M. Hemalatha [5]detected the severity of attacks in the dataset based on kddcup99 dataset produced by MIT Lincoln Laboratory.
Marek Kretowski.Marek Gizes, Bialystok Technical University, Poland [6] have presented a new evolutionary algorithm (EA) for induction of mixed decision tree.[6] In non-terminal nodes of a mixed tree, different types of tests can BE placed, ranging from typical inequality test up to an oblique test based on a splitting hyperactive plane. In contrast to classical top down methods, the proposed system searches for an optimal tree in a global manner that is it learns a tree structure and finds tests in one run of EA . Specialized genetic operators are developed, which allow the system to exchange parts of trees, generating new sub trees, pruning existing one and changing the node type and the test. An informed mutation application scheme introduced and the number of unprofitable modification reduced. All the works were using different algorithms for prediction.

3. METHODOLOGY

Data mining can take on different approaches and build different models depending upon the type of data involved and the objectives. This research work is based on different data mining algorithms on multi dimensional data analysis. . The common models used in predictive data mining includes Association rules,Clustering methods,Decision tree,Classification Rules and Statistical mining tools.
For disease identification and prediction in data mining a new hybrid algorithm was constructed. The Disease Identification and Prediction(DIP) algorithm, which is the combination of decision tree and association rule is used to predict the chances of some of the disease hits in some particular areas. It also shows the relationship between different parameters for the prediction. To implement this algorithm using vb.net software was also developed.

At the conceptual level, using Decision tree create a non-linear data mining model for correlating variables. By traversing the Decision tree from root to leaves the prediction rules is directly obtained. The logical dependency between various attributes of an entity using association rule of Apriori principle is constructed . By measuring the confidence and support, association strengths can be measured. Using a minimum confidence and support thresholds, the rule-mining algorithm identifies all association satisfying the specified parameters and find the dependencies between different attributes of the same entity. We can apply the rules and data to a statistical technique in the cluster analysis to extract all possible clusters from unlabelled data. The results can be represented in graphical form for analysis.

3.1 Phase 1:

Create a non-linear DM model for co-relating parameters using DT and develop the prediction rule. Let T be the training data set with class labels{c1, c2,….ck} and X is the non-class attributes of T. Form the attribute list of X w.r.t T and sorted the attribute list and using relevance analysis made attribute removal using the threshold.We can Measure uncertainty coefficient of an attribute X using the equation, UC(x)=gain(x,T)/info(T) and Gain(x,T)=info(T)-info(x,T).

3.2 Phase 2:

Find the frequent items and store in the compact structure. Merge the sets and registered as a count if multiple transaction. Scan the database and find all the frequent items and their support. Create the tree with root as null. Get the first transaction from the database, remove all non-frequent items, and list the remaining according to the order in sorted frequent items. Use the transaction to construct the first branch of the tree with each node corresponds to frequent item. Get the next transaction, remove non-frequent items, insert in the tree, and increase item count .Continue until all transactions are completed.

4. ANALYSIS REPORTS

Fig:1 Implementation screen

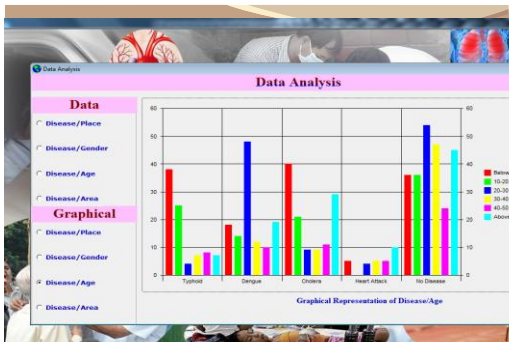

Fig:2 Disease-age representation

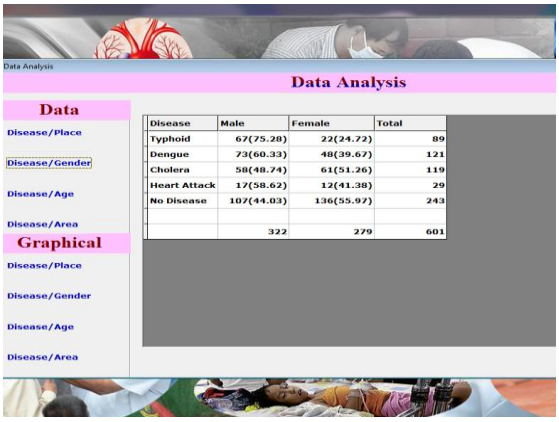

Fig.3: Data Analysis screen

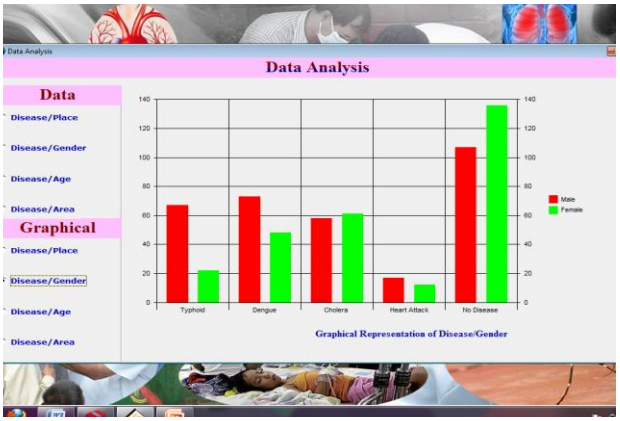

Fig 4: disease-gender representation.

4. 2. IMPLEMENTATION AND RESULTS

Table 1: Results obtained from DIP

Disease	Area	Gender	Age group	Sanitation	Food source	water source	type of food	Hereditary	Income	Nature of job
Typhoid	Urban	Male	0-10, 11-20	Average, poor	Outside	Open source	Non Vegetarian	No	NA	NA, self emp
Dengue	Urban	Male	21-30	Good	Home	NA	Non-Vegetarian		NA	NA, self emp
Cholera	Rural	Female	0-10	Average, poor	Outside	Open source	Non-Vegetarian	No	NA	NA, self emp
Heart Attack	Urban	Male	40 above	Good	Outside	NA	Non-Vegetarian	Yes	High salaried	self emp

5. CONCLUSION

The proposed new hybrid algorithm is unique and different from the commonly used prediction algorithms in data mining. The proposed methods overcome the disadvantages of existing methods as the number of frequent items is less. The new algorithm proved efficient in terms of time and space complexity and proved to be accurate when compared with a standard statistical analysis tool such as SPSS.

6. FUTURE ENHANCEMENT

This hybrid algorithm can be enhanced by considering and incorporating many more parameters in the cluster. For disease identification and prediction for agricultural diseases, the same set of algorithms and rules can also apply.

7. ACKNOWLEDGEMENT

I express my sincere gratitude to God Almighty for all his blessings showered upon me for the completion of this research work. I am heartily thankful to my supervisor, Dr. V. Sundaram, whose encouragement, guidance, supervision, and support from the initial to the final level enabled me to complete the research work.

8. REFERENCES

[1]. Arijay Chaudhry and Dr. P.S.Deshpande. Multidimensional Data Analysis and data mining, Black Book

[2]. Smitha.T ,Dr.V.Sundaram"Case study on High Dimensional Data Analysis using Decision Tree model", , International journal of computer science issues Vol9,Issue 3, May 2012.

[3] Smitha.T,Dr.V.Sundaram"Classification Rules By Decision Tree for disease Prediction", ,International journal of Computer Applications vol-43, No-8, April 2012.

[4]. "Smitha.T, Dr.V.Sundaram" Knowledge Discovery from Real Time Database using Data Mining Technique, IJSRP vol 2, issue 4, April 2012.

[5]. Moawia Elfaki Yahia1, Murtada El-mukashfi El-taher2 "A New Approach for Evaluation of Data Mining Techniques", ,IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.

[6].V.Umarani "A study on effective mining of association rules from huge database" al. / IJCSR International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010.

[7]. Shalini S Singh " K-means v/s K-medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, May 2011.

[8].C. MÁRQUEZ-VERA" Predicting School Failure Using Data Mining" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010

[9].Smitha.T, Dr.V.Sundaram "Comparative study of data mining algorithm for high dimensional data analysis" International journal of advances in Engineering & Technology, Vol 4, issue 2, ISSN. 2231-1963, Sept-12, pp. 173-178.

[10].G.SenthilKumar "online message categorization using Apriori algorithm" International Journal of Computer Trends and Technology- May to June Issue 2011.

[11] Behrouz Minaei-Bidgoli, Elham "A New Approach of Using Association Rule Mining in Customer Complaint Management" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010

[12] .K.Srinivas et al. "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010

[13] Sunita Soni, Jyoti Soni, Ujma Ansari "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" *International Journal of Computer Applications (0975 – 8887) Volume 17– No.8.*

[14] .Shelly Gupta et al. "DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS " Indian Journal of Computer Science and Engineering (IJCSE) March 2011.

AUTHOR'S PROFILE

First Author: Mrs.Smitha.T, obtained her post graduate degree in Computer Applications and M.Phil in computer science from MK university ..Doing PhD in computer science at Karpagam

Unicersity, Coimbatore. She is now working as an Asst.Professor in the Dept of computer Applications, Sree Narayana Guru Institute of Science and Technology, Paravoor, Kerala. She is interested in data mining studies and its applications. She has published 5 different papers in international journals and presented many papers in international and national seminars regarding data mining and warehousing.

Second Author: Dr.V.Sundaram

Dr.V.Sundaram. He acquired his post graduation in Mathematics and PhD in applied mathematics. He has 45 years of teaching experience in India and abroad and guiding more than 10 scholars in PhD and M.phil at Karpagam and Anna University. He has organized and presented more than 40 papers in national as well as international conferences and have many publications in international and national journals. He was the former Director of MCA Department of Karpagam Engineering College. He is a life member in many associations.His area of specialization includes fluid Mechanics, Applied athematics, Theoretical Computer Science, Data mining, and Networking and security etc.