Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A hybrid Algorithm for Epidemic Disease Prediction with Multi Dimensional Data

N K Kameswara Rao[*],
Assoc. Professor,
Department of Information Technology,
SRKR Engineering College,
Bhimavaram, Andhra Pradesh, India

Dr. G P Saradhi Varma
Professor & HOD,
Department of Information Technology
SRKR Engineering College,
Bhimavaram, Andhra Pradesh, India

Abstract— Data mining is has three major components Clustering or Classification, Association Rules and Sequence Analysis. The clustering techniques analyze a set of data and generate a set of grouping rules that can be used to classify future data. The mining tool automatically identifies the clusters, by studying the pattern in the training data. Once the clusters are generated, classification can be used to identify, to which particular cluster, an input belongs. The main aim this paper is to discover the a fast, easy and an efficient data mining algorithm in the prediction of Epidemic Disease in an area with very lees number of error and also can work with very large sets of data and show reasonable patterns with dependent variables. For prediction and identification of Epidemic Disease in data mining we proposed as new hybrid algorithm, Epidemic Disease Prediction and Identification (EDPI) algorithm. This is a combination of decision tree and association rule mining to predict the changes of getting Epidemic Disease in some selected areas. The prediction of disease in this algorithm can be shows by the relationships between desired parameters. The implementation of the algorithm is developed in Visual basic language.

Keywords—Association rule, Classification rule, Clustering, Data mining, Decision Tree, Multidimensional data, Prediction.

I. INTRODUCTION

The automatic extraction of implicit and interesting patterns from large data collections form the data base is called Data mining. The process of discovering patterns in large data sets is called data mining and it is the combination of statistics and computer science. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics and database systems. Extraction of information from a data sets and transforms it into an understandable structure for future use is the overall goal of the data mining process [1]. Data mining is a system which searching for patterns in large amounts of data. The main goal of data mining is previously unknown important information can be extracted from data. It is commonly used to identify 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. More number of Problems can be solved by using the above found patterns [9].

The people who are using data mining can be able to predict certain behaviors or patterns. Once the user is able to predict the behavior of something he can analyze and will be able to make strategic decisions that can allow him to achieve certain goals.

II. RELATED WORK

The diagnosis of the disease as exact as possible and helps to make a decision if it is reasonable to start the treatment on suspected patients without waiting for the precise medical test result or not by developing a data mining solution is the main purpose the this study. Identification of different parameters and the analization of the facts and reasons behind the disease are also considered in this study. Proper data for disease must be identified, preprocessed, transformed and loaded into data warehouse system, store and manage the data into a multidimensional database system can be extracted by data mining. The data can be viewed in a presentable format is provided to the data access by an analyst.

III. 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.

IV. DATA COLLECTION:

To find the prediction of infectious disease hit, different kinds of data were collected from different sources 8 different areas are selected contains two urban, two rural and two tribal areas and 2 hill areas in East Godavari District in Andhra Pradesh. 2424 records are collected from the above 8 areas.

After each rainy season, some contagious disease is hitting in many families in these areas. This is repeating since the last several years. So, all the data have been collected from atleast two from each family. The main parameters were education, income, hereditary factors, area located, drainage facilities, drinking water facilities, toilet facilities, waste disposal, electricity, approaches to hospital, roads, educational institutions, livelihood etc and created a database. Based on different parameters classified the population into different levels and created suitable model based on the selected algorithm [11].

The raw data used in the research were collected from health department, Hospital, Urban Local Body, inhabitants from the above areas, Doctors from various hospital, health officers, different records from urban local body, on site observation etc. The same type of data collected from inhabitants inside the slum as well as outside the slum. To ensure the consistency of result, missing values were also dealt with. Irreverent records and duplicated data were eliminated to reduce the size of data set. Data synchronization was also carried out.

V. METHODOLOGY:

Data mining can take on different approaches and build different models depending upon the type of data involved and the objectives. Different data mining algorithms on multi dimensional data analysis is used in this research work. Association rules, Clustering methods, Decision tree, Classification Rules and Statistical mining tools are used as common models for prediction in data mining.

For disease prediction and identification in data mining a new hybrid algorithm was constructed. The Epidemic Disease Prediction and Identification (EDPI) algorithm, This is a combination of decision tree and association rule mining to predict the changes of getting Epidemic Disease in some selected areas. The prediction of disease in this algorithm can be shows by the relationships between desired parameters. The implementation of the algorithm is developed in Visual basic language.

At the theoretical level, we created a non-linear data mining model for correlating variables b using Decision tree algorithm. 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. Association strengths are measured by measuring the confidence and support. 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. The implementation of this was done by 2 phases.

V-I Phase 1:

Created a non-linear Data Mining model by using Decision tree and developed a prediction rule for co-relating parameters. 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 Measured 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).

V-II Phase 2:

Find the frequent items and stored in the compact structure. Merged the sets and registered as a count if multiple transaction. Find all the frequent items and their support by scanned the database. A tree was created with a root node as null. Removed all non-frequent items, and listed the remaining according to the order in sorted frequent items depending upon the first transaction from the database. Used the transaction to construct the first branch of the tree with each node corresponds to frequent item. Removed non-frequent items and inserted in the tree, and increased item count until all transactions were completed depending upon the next transactions.

VI. ANALYSIS REPORTS:

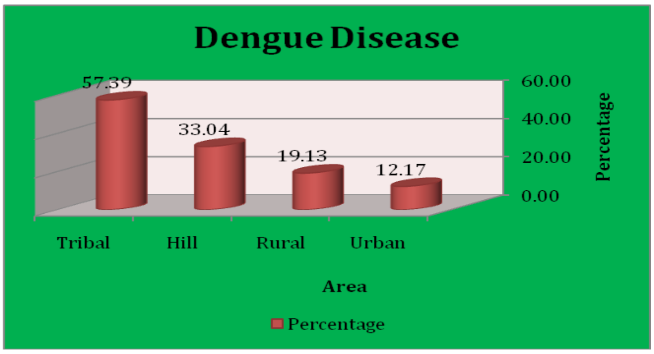

Fig-1 People effected dengue disaster Area-wise

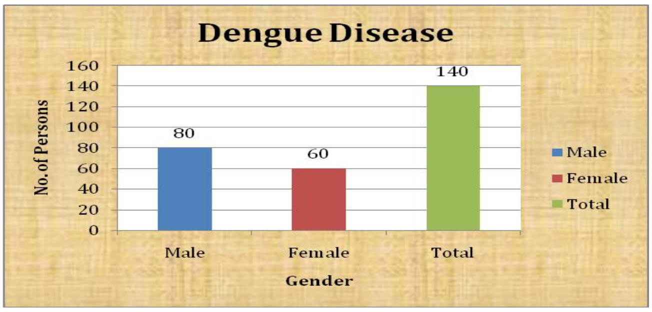

Fig-2 People effected dengue disaster Gender-wise

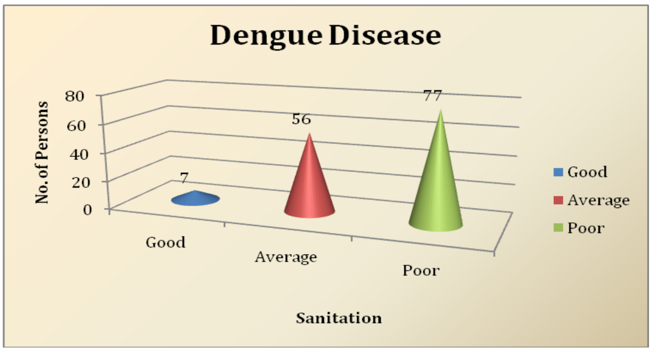

Fig -3 People effected dengue disaster Sanitation-wise

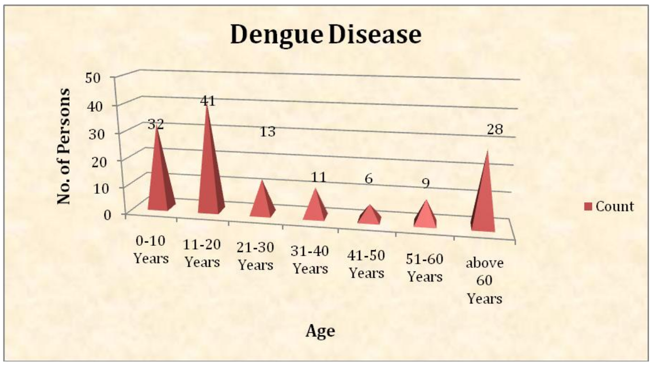

Fig -4 People effected dengue disaster Age Group-wise

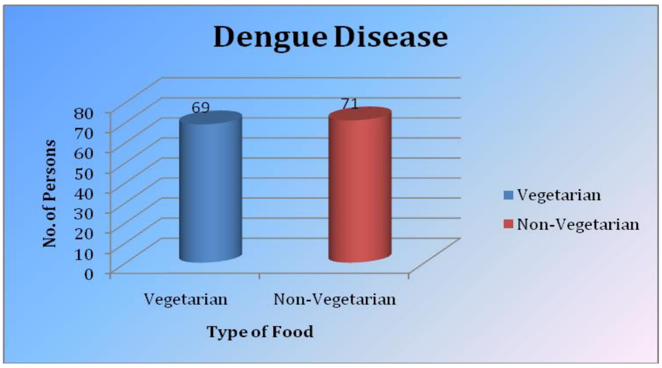

Fig -5 People effected dengue disaster Type of food-wise

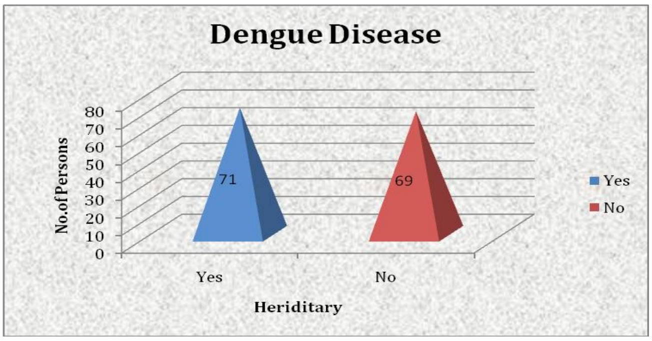

Fig -6 People effected dengue disaster with Hereditary

VI-I. IMPLEMENTATION AND RESULTS

TABLE 1: RESULTS OBTAINED FROM EDPI

Disease	**Area**	**Gender**	**Age group**	**Sanitation**	**Food source**	**water source**	**type of food**
Dengue	Tribal	Male	0-10, 11-20	Good	Outside	Open source	Vegetarian
Dengue	Hill	Female	21-30	Average, poor	Home	Open source	Vegetarian
Dengue	Rural	Female	0-10	Average, poor	Outside	Public	Non-Vegetarian
Dengue	Urban	Male	above 60	Poor	Outside	Public	Non-Vegetarian

VII. 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 weka tool.

VIII. 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.

Acknowledgment

Sincere thanks to management, Principal and faculty members of our institute (Sagi Ramakrishnam Raju Engineering College).

REFERENCES

[1] Jaiwei Han; Micheline kamber; Data mining concepts and Techniques; Morgan Kaufmann Publishers.

[2] Fayyad U.M.Piatetsky-Shapiro.G & smith.P" From data mining to knowledge discovery in databases' AI magazine 17(3) pp-37-54.

[3] Ms.Sunu Mary Abraham"User Behaviour Based Clustering and Decision Tree Model for predicting customer insolvency in Telecommunication Business. Karpagam Journal-Jan-2011, Volume 5

[4] K.S.Adekeye and M.A.Lamidi, "Prediction Intervals: A tool for monitoring outbreak of diseases" International journal for data Analysis and information System jan-2011-Vol-3.

[5] Aitchison.J and Dunsmore, Statistical Prediction Analysis: Cambridge University Press.

[6] Waleed Alsabhan and Oualid Ben Ali " A new multimodal approach using data mining: the case of jobseekers in the USA" International journal for data Analysis and information System jan-2011-Vol-3.

[7] Rui Xu , Donald C.and Wunsch Clustering, Iee Press-2008.

[8] Bori Mirkin(2005) clustering for Data mining Chapman & Hall/Crc.

[9] Apte, C.and Weiss,S.M(1997), " Data mining with Decision Trees and Decision Rules" Future generation computer systems, 13,197-210.

[10] Ch.Ding, X.He"K means clustering via principal component Analysis Proc.of international conference on machine learning(2004),pp.225-232,2004.

[11] N K Kameswara Rao and G P. Saradhi Varma "Classification Rules Using Decision Tree for Dengue Disease " IJRCCT, Vol 3, issue 3, March 2014.

AUTHORS:

N K Kameswara Rao did M Tech in Software Engineering from JNT University-Hyderabad. Presently he is working as Associate Professor in Information Technology Department, SRKREC – Bhimavaram.

Dr. G P. Saradhi Varma did Ph. D in Computer Science & Systems Engineering from Andhra University-Visakhapatnam. Presently he is working as Director PG Courses, Professor and HOD in Information Technology Department, SRKREC – Bhimavaram. He Guided 15 Ph D. scholars. He published 26 National Journals, 37 International Journals and 6 books.