ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

EFFICIENT MODEL FOR CHD USING ASSOCIATION RULE WITH FDS

Priyanka Palod[1], Jayesh Gangrade[2]

P.G. Student, Department of Computer Engineering, I.E.S. IPS ACADEMY, A.B. ROAD, INDORE, India[1]

Associate Professor, Department of Computer Engineering, I.E.S. IPS ACADEMY, A.B. ROAD, INDORE, India[2]

Abstract: Association rules are an energetic investigating area. Association rules characterize a promising method to search syndrome differentiation on modern India. Solitary of the most accepted approach to do data mining is determining association rules. The association innovation is an imperative research field in data mining. The mining association rule frequently has been adopts numerous models: support, confidence, interestingness. But this model can't accurate measure the correlative degree between the precursor and the consequential of the rule by allocation. So we proposed a new mining model of association rules: support, coincidence, interestingness and investigate the significance of fluke by instance. We use this model in the data about coronary heart disease and obtained a lot of meaningful rules. Proposed a new model of support-coincidence-interestingness base on the traditional model of support-confidence interestingness. Our propose model can quantitatively evaluate the correlation of rules and reduce many rules that have low support or have no correlation or have negative correlation. In our work we will conduct experiments on large real time to predict the diseases like Medication in Coronary Heart Disease and compare the performance of our algorithm with other related algorithms. Our propose model based on CMAR (Classification based on Multiple Association Rules) SVM, fuzzy discriminant Analysis.

Kew words: Coronary Heart Disease, SVM, fuzzy Discriminant fuzzy diminishing support.

I. INTRODUCTION

With the ever-growing complexity in recent years, huge amounts of information in the area of medicine have been saved every day in different electronic forms such as Electronic Health Records (EHRs) and registers. These data are collected and used for different purposes. Data stored in registers are used mainly for monitoring and analyzing health and social conditions in the population. The unique personal identification number of every inhabitant enables linkage of exposure and outcome data spanning several decades and obtained from different sources. The existence of accurate epidemiological registers a basic prerequisite for monitoring and analyzing health and social conditions in the population. Some registers are state-wide, cover the whole collieries population, and have been collecting data for decades. They are frequently used for research, evaluation, planning and other purposes by a variety of users in terms of analyzing and predicting the health status of individuals. Regardless of the research activities to prevent Coronary Heart Disease (CHD), it vestiges one of the most imperative reason of death in India. Coronary Heart Disease (CHD) are frequently deadly and most of the people, who die because of it, have experienced different symptoms that were not taken into account. The facts illustrate that, each year, almost 100000 people die since of Coronary Heart Disease. Clustering method are extensively used for non supervised learning since they are able to separate a finite unlabeled data set into finite sets. They are commonly used in different areas, but in medical research, several works are related to the identification of different sufferings or pathologies. In our propose technique, one obtains initially fuzzy Discriminate Analysis, new association rules which are either used in the model directly or after projection onto the individual antecedent variables. The performance of the algorithms depend on the number of clusters in data, the shape and volume of every cluster, the initialization of the clustering algorithm and the distribution of the data patterns. Thus, in order to improve their performance, it is needed to combine them with other algorithms like Support Vector Machines (SVM). We present a new association rule mining Model: fuzzy decreasing support-confidence that finds all item sets that satisfy a length-decreasing support constraint. On this basis, by analyzing the correlation between the antecedent and the consequent of the generated rules, Bidirectional

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

Confidence, Interestingness; We mining data regarding the applicable aspect of disease favoritism and the patients' medication from the coronary heart disease data collected from the hospitals. The investigational consequences demonstrate that the model propose in this research not merely verify the existing Syndrome Differentiation and regular patterns of medication, but also discover Syndrome Differentiation with a combination of factors and medicine compatibilities among multiple drugs. We chose to combine mining association rules ,fuzzy Discriminant Analysis with the SVM algorithm because this learning method provides a convergence to a globally optimal clarification and for several problems it has exposed better generalization capabilities than other learning techniques.

a) Data Mining concepts in Health Care

Data Mining aims at discovering knowledge out of data and presenting it in a form that is easily compressible to humans. It is a process that is developed to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. In practice, the two primary goals of data mining tend to be prediction and description [1, 2]. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans Basic concepts and terminology

b) Association Rules

Formally, association rules are defined as follows: Let $I = \{i_1, i_2 \ldots i_n\}$ be a set of items, D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with a unique identifier TID. A transaction T is said to contain X, a set of items in I, if $X \subseteq T$. An association rule is an implication of the form "$X \rightarrow Y$", where $X \subseteq I$; $Y \subseteq I$, and $X \cap Y = \Phi$. The rule $X \rightarrow Y$ has support s in the transaction set D if s% of the transactions in D contains X U Y. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \rightarrow Y$ holds in the transaction set D with confidence c. If c% of transactions in D that contain X also contain Y. In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a support and confidence greater than given thresholds. These rules are called strong rules, and the framework is known as the support-confidence framework for association rule mining Transforming

c) Medical Data Set

A medical dataset with numeric and categorical attributes must be transformed to binary dimensions, in order to use association rules. Numeric attributes are binned into intervals and each interval is mapped to an item. Categorical attributes are transformed by mapping each categorical value to one item. Our first constraint is the negation of an attribute, which makes search more exhaustive. If an attribute has negation then additional items are created corresponding to each negated categorical value or each negated interval. Missing values are assigned to additional items, but they are not used. In short, each transaction is a set of items and each item corresponds to the presence or absence of one categorical value or one numeric interval.

II. SVM (SUPPORT VECTOR MACHINE)

A SVM algorithm for the classification of both linear and nonlinear data. It transforms the original data in a higher dimension, from where it can find a hyper-plane for separation of the data using essential training examples called support vectors**.** The SVM is a basically two class classifier and can be extended for multi-class classification. In our model we use each object is mapped to a point in a high dimensional space, each dimension of which corresponds to features. The coordinates of the point are the frequencies of the features in the corresponding dimensions. SVM learns, in the training step, the maximum-margin hyper-planes separating each class. CPAR (Classification based on Predictive Association Rules). CMAR adopts method of frequent item set mining to generate candidate rules. However, CMAR generates a large number of rules. CMAR (Classification based on Multiple Association Rules). CMAR is the associative classification. CMAR generates rules using the FP-growth algorithm. In the pruning phase, CMAR selects only positively correlated rules

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

III. BRIEF LITERATURE SURVEY

We start by discussing research on data mining techniques used with medical and biological data. Important issues [1] when using machine learning or data mining techniques in the medical domain include fragmented data collection, strict privacy regulations, rich data types (image, numeric, categorical, missing information), complex taxonomies classifying attributes, and an already rich and complex knowledge base.

Zheng-kui Lin at al[4] The mining association rule usually adopts this model: support, confidence, interestingness. But this model can't measure the correlative degree between the antecedent and the consequent of the rule by ration. So they proposed a new mining model of association rules: support, coincidence, interestingness and analyzed the meaning of coincidence by instance. At last, they used this model in the data about coronary heart disease and obtained a lot of meaningful rules.

Jyoti Soni at al [5] the healthcare environment is still „information rich but „knowledge poor. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research intends to provide a of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome 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.

Himigiri. Danapana in at al[6] This research intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of decision tree.

Fariba Shadabi in at al[7] Artificially Intelligent (AI) powered tools are able to deal with uncertain and incomplete data sets. Neural network classifiers have been successfully used for prediction purposes in many complex situations. Research demonstrates that AI-based data mining tools have been also successfully used in many medical environments. This research advances the understanding of the application of Artificial Intelligence and Data Mining tools to clinical data by demonstrating the potential of these techniques in complex clinical situations.

Sunil Joshi in at al [8] recently, different works proposed a new way to mine patterns in transposed databases where a database with thousands of attributes but only tens of objects. In this case, mining the transposed database runs through a smaller search space. In this research, they systematically explore the search space of frequent patterns mining and represent database in transposed form. They develop an algorithm (termed DFPMT—A Dynamic Approach for Frequent Patterns Mining Using Transposition of Database) for mining frequent patterns which are based on Apriori algorithm and used Dynamic Approach like Longest Common Subsequence. The main distinguishing factors among the proposed schemes is the database stores in transposed form and in each iteration database is filter /reduce by generating LCS of transaction id for each pattern. Their solutions provide faster result. A quantitative exploration of these tradeoffs is conducted through an extensive experimental study on synthetic and real-life data sets.

T.John Peter in at al [9] in this research, the use of pattern recognition and data mining techniques into risk prediction models in the clinical domain of cardiovascular medicine is proposed. The data is to be modeled and classified by using classification data mining technique. Some of the limitations of the conventional medical scoring systems are that there is a presence of intrinsic linear combinations of variables in the input set and hence they are not adept at modeling nonlinear complex interactions in medical domains. This limitation is handled in this research by use of classification models which can implicitly detect complex nonlinear relationships between dependent and independent variables as well as the ability to detect all possible interactions between predictor variables.

Rahul Isola in at al [11] The system proposed in this research uses this vast storage of information so that diagnosis based on this historical data can be made. It focuses on computing the probability of occurrence of a particular ailment from the medical data by mining it using a unique algorithm which increases accuracy of such diagnosis by combining the key points of Neural Networks, Large Memory Storage and Retrieval (LAMSTAR), k-NN and Differential Diagnosis all integrated into one single algorithm.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

IV. PROPOSED METHODOLOGY

The system uses a Service-Oriented Architecture wherein the system components of diagnosis, information portal and other miscellaneous services are provided. This algorithm can be used in solving a few common problems that are encountered in automated diagnosis these days, which include: diagnosis of multiple diseases showing similar symptoms, diagnosis of a person suffering from multiple diseases, receiving faster and more accurate second opinion and faster identification of trends present in the medical records. Coronary Heart Disease (CHD) is a common cardiovascular disease which does seriously harm to humans' health. It belongs to the category of chest stuffiness and pains on Traditional Indian Medicine (TIM). The research of TIM syndrome differentiation could help to improve the level of TIM treatment based on syndrome differentiation and clinical medication programs. Currently, the main research of CHD differentiation [1-3] involves in coronary angiography, QT dispersion and JT dispersion, ultrasonic cardiogram, blood lipid and so on. Mining association rules are absorbed into searching syndrome differentiation of CHD.

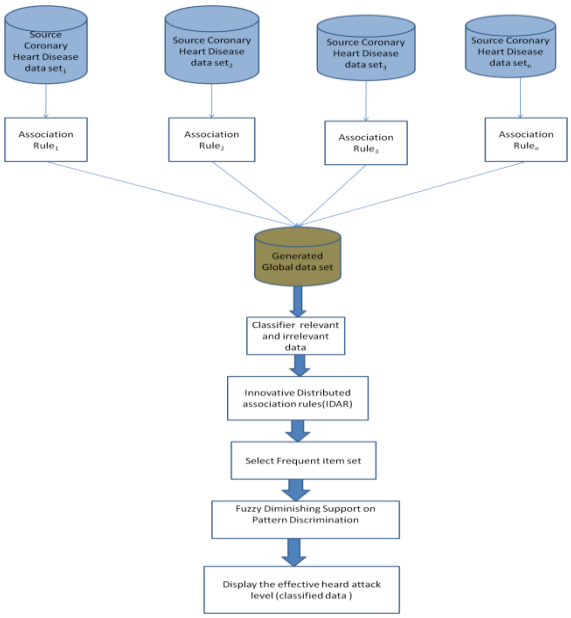

Figure 1: System architecture

.

V. HEART DISEASE DATASET

Attribute
Age
Smoking
Over weight
resting blood pressure
serum cholestoral in mg/dl
fasting blood sugar > 120 mg/dl
Alchol intake
maximum heart rate achieved
exercise induced angina
the slope of the peak exercise ST segment
number of major vessels (0-3) colored by flourosopy
Hereditary
diagnosis of heart disease

Data oriented study becomes a new way for TCM syndrome differentiation of modern CHD. By analyzing the syndrome differentiation experiences from the famous or old traditional Chinese medicine doctors, we can discover some laws of syndrome differentiation and improve the accuracy of differentiation of doctors' clinical CHD.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

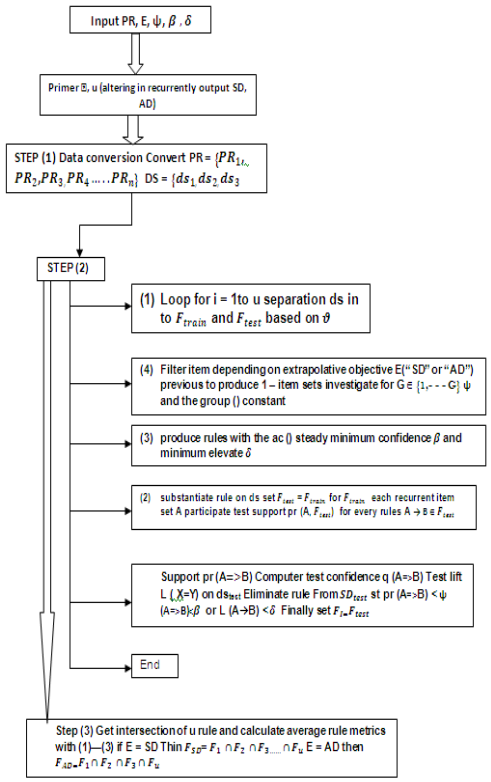

Figure 2: Algorithm for heart disease prediction

Therefore, the introduction of data mining methods into the syndrome differentiation of CHD will have a huge theoretical and practical significance. According to the characteristics of TCM data, mining association rules based on the attribute constraint and the antecedent-consequent constraint is studied in [4]. In [5] the author uses search constraint to discover interesting association rules and reduces the number of generated rules. However, these approaches above make improvements on the basis of a constant support threshold, so they still can not mine the effective long patterns. According to the characteristics of CHD data, this paper uses the association rule mining algorithm with fuzzy decreasing support constraint to mine the relative factors of differentiation.

VI. CONCLUSION

We used association rules to predict the degree based on heart perfusion measurements and risk factors. We studied two complementary tasks: predicting the absence and predicting the existence of heart disease. We focused on two main research issues. The first issue is the large number of rules that are obtained by the standard association rule algorithm. The second issue is the validation of rules on an independent set, which is required to eliminate unreliable rules, or rules that cannot be generalized. Constraints were proposed in this work to reduce the number of rules: item filtering, attribute grouping, maximum item set size, and antecedent/consequent rule filtering. Contrasting with previous work, our constraints are specified on raw attributes instead of items, item filtering is applied earlier before generating frequent item sets, and the group constraint induces a partition on attributes allowing easier manipulation. In order to validate rules, we used the train and test approach that uses two disjoint samples from a data set to search and validate rules. All features are assembled together in one algorithm that combines search constraints and train/test validation. The algorithm performs several train and test cycles to achieve basic cross-validation and reduce the number of rules with poor generalization potential.

REFERENCES

[1]Carlos Ordonez and Kai Zhao,"Evaluating association rules and decision trees to predict multiple target attributes", Intelligent data Analysis 15 (2011) 173–192 173,DOI 10.3233/IDA20100462,IOS Press.-2011
[2] Mannila, H.: Methods and Problems in Data Mining. In: The International Conference on Database Theory, pp. 41–55 (1997)
[3] J. F. Roddick, P. Fule, and W. J. Graco, "Exploratory medical knowledge discovery: Experiences and issues," SIGKDD Explorations, vol. 5, no. 1, pp. 94–99, 2003.
[4]Zheng-kui Lin1,Wei-guo Yi1, Ming-yu Lu1, Hao Xu2 Zhi Liu1Correlation Research of Association Rules and Application in the Data about Coronary Heart Disease- 978-0-7695-3879-2/09-IEEE-2009.
[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni" Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.
[6]1Himigiri. Danapana, M. Sumender Roy," Effective Data Mining Association Rules for Heart Disease Prediction System" IJCST Vol. 2, Issue 4, Oct . - Dec. 2011.
[7]Fariba Shadabi and Dharmendra Sharma," Artificial Intelligence and Data Mining Techniques in Medicine – Success Stories" International Conference on BioMedical Engineering and Informatics- 2008.
[8] Sunil Joshi, Dr. R. C. Jain," A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database" Second International Conference on Communication Software and Networks- 2010.
[9]T.john peter, k. somasundaram, an empirical study on prediction of heart disease using classification data mining techniques-IEEE-international conference on advances in engineering, science and management (icaesm -2012) march 30, 31, 2012.
[10] zhe wang, mingsan miao," discovery the relationship in properties of traditional chinese medicine
based on data mining" international symposium on information technology in medicine and education-2012.
[11] Rahul isola, rebeck carvalho and amiya kumar tripathy." knowledge discovery in medical systems using differential diagnosis, lampstar and k-nn" IEEE transactions on information technology in biomedicine - titb-00346-2011.