



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Performance of Classification in Medical Data Mining

Dr.K.Ravichandran¹, S.Nagarasan²

Assistant Professor & Head, Dept. of Computer Application, H.H the Rajah's College (Autonomous), Pudukkottai,
India¹

Research Scholar, H.H the Rajah's College (Autonomous), Pudukkottai, India²

ABSTRACT: The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. A feature selection is an important step in classification and also for dimensionality reduction. As medical information is with multiple attributes, medical data mining differs from other one. Diagnosis of most of the diseases is expensive as many tests are required to predict the disease. By using data mining techniques we can reduce the cost of diagnosis by avoiding many tests by selection of those attributes which are really important for prediction of disease. Dimensionality reduction plays an important role in the field of medicine as it contains multiple attributes. In this paper we have analyzed the approach of feature selection for classification and also presented a novel approach for the feature selection by using association and correlation mechanism. The aim of our paper is to select the correlated features or attributes of medical dataset so that patient need not to go for many tests and in future it is used for preparing the clinical decision support system which is helpful for decision making of disease prediction in a cheaper way. Other approach is mentioned in this paper is after removal of some attributes accuracy of classifier is also improved which support our statement of disease prediction in cheaper way by avoiding all unwanted tests for disease prediction. By using association rules and correlation attributes features can be selected. As medical field contains large number of attributes and information so dimensionality reduction is must now. The accuracy of classifiers after removal of attributes is discussed in this paper.

KEYWORDS: Classification and association, Data mining, Heart diseases, Medical data mining, Knowledge Discovery.

I. INTRODUCTION

In present days, computers have brought significant improvements to technology that lead to the creation of huge volumes of data. Moreover, the advancement of the healthcare database management systems creates a huge number of medical databases. Creating knowledge and management of large amounts of heterogeneous data has become a major field of research, namely data mining. Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. Using training dataset we trained the classifier. Correctness of the classifier could be tested using test dataset. In the present study, we have focused on the usage of classification techniques in the field of medical science and bioinformatics. Classification is the most commonly applied data mining technique, and employs a set of pre-classified examples to develop a model that can classify the population of records at large. The major goal of the classification technique is to predict the target class accurately for each case in the data. This is also called knowledge data discovery is the process of analyzing data from different perspective and summarizing it into useful information. Knowledge Discovery in database is concerned with the development of methods and techniques for making sense of data. KDD is the process of mapping low-level data into other forms that might be more compact, more abstract or more useful. Main aim of data mining is to uncover

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

relationship in data and predict the outcome. [1] Data mining extracts the patterns in the process of knowledge discovery in the database. As the dataset has grown in size and complexity, new emerging fields of data mining provide new techniques and methods which help to analyze and understand large bodies of data. Data mining involves some common methods that are association rule learning, clustering, classification, regression, summarization and sequential pattern matching. Classification is one of the important techniques of data mining. Classification is the process of finding a set of models (or functions) which describe and distinguish data classes or concepts [2]. In classification, inputs are given a set of data, called a training set, where each record consists of several fields or attributes. These attributes are continuous, coming from an ordered domain, or categorical, coming from an unordered domain. One of the attributes, called the classifying attribute, indicates the class to which each dataset belongs. The objective of classification is the method to build a model of the classifying attribute based upon the other attributes which are not from the training data set. Data Mining techniques are also used to analyze the various factors that are responsible for diseases for example type of food, different working environment, education level, living conditions, availability of pure water, health care services, cultural, environmental and agricultural factors.

Attributes	Values
Age	Numerical
Sex	Male, female
Chest pain type	1, 2, 3, 4
Resting blood pressure	Numerical
Serum cholesterol in mg dL^{-1}	Numerical
Fasting blood sugar $>120 \text{ mg d}^{-1}$	Yes, no
Resting electrocardiographic results	0, 1, 2, 3
Maximum heart rate achieved	Numerical
Exercise induced angina	Yes, no
ST depression induced by exercise relative to test	Numerical
The slope the peak exercise ST segment	Numerical
Number of major vessels colored by fluoroscopy	0, 1, 2, 3
Thal	Normal, fixed defect, Reversible defect
Absence or presence of heart disease	Absence, presence

Figure 1: number of attributes and number of values

II. RELATED WORK

Medical Data Mining: Hanauer [2] reported the challenges and solutions in mining electronic data for research and patient care. The Michigan Health system statistics were utilized for their research. However the author was concerned and focused on the hurdles involved in text mining alone. The challenges that the author inferred included affirmation of accurate diagnosis and natural language processing of electronic health records. The author had provided a solution called EMERSE (Electronic Medical Record Search Engine) that provided keyword searches for basic users and advanced features for power users. The interface was user-friendly, secure and compliant with privacy regulations and practical for implementation. However the system needed more training and the searching procedures continued to raise complexity. Roddick et.al, [10] presented the experiences of the authors in applying exploratory data mining techniques to medical health and clinical data. This enabled the authors to elicit a number of general issues and provided pointers to possible areas of future research in data mining and knowledge discovery from a broad perspective. Iavindrasana et.al [7], used the nine data mining steps proposed by Fayyad in 1996 [8] as the main themes of the review. MEDLINE [6] was used as the primary source and 84 papers were retained by the authors for analysis. Their results identified three main objectives of data mining that were stated as follows: understanding of the clinical data, providing assistance to healthcare professionals, and formulating a data analysis methodology to explore clinical data. Classification was stated to be the most frequently used data mining function with a predominance of the implementation of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines). A myriad of quantitative performance measures were proposed with a predominance of accuracy, sensitivity, specificity, and ROC curves. Further work was reported by Lalayants et.al [4] who described



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

a practice-based, mixed-method research methodology stating Clinical Data Mining (CDM) to be a strategy for engaging international practitioners for describing, evaluating and ruminating upon endogenous forms of practice with the ultimate goal of improving practice and contributing to knowledge[9]. Such knowledge contributions were considered to be localized, but through conceptual reflection with empirical replication they could be generalized. Data Mining Models in CDM Clinical data mining analysis crafts effective and worthwhile knowledge that is indispensable for precise and accurate decision making [2]. Various types of mining models have been used in the past to represent interesting facts and latent patterns and trends in clinical datasets with copious applications in medical practice [2] [3]. In this subsection some of the data mining models applied to healthcare are briefly reviewed. Feature Relevance Models Clinical data are generally voluminous in nature and need special attention by virtue of data storage and analysis. Feature relevance analysis[4][5] is a phase in data mining that enables researchers to filter out certain predictors of ailments from further exploration under the pretext of being less contributory to the detection of an ailment[26]. For instance, a patient's health record may contain the concerned Patient ID, Address, and Occupation along with the evidenced clinical findings and laboratory investigation results among other details. The former factors are highly inessential in diagnosing the patient's state of health and time spent on analysis of such details is a huge squander. Such attributes need to be filtered out from further analysis and this would certainly save time and lessen computational complexity. Clustering Models Clustering is derived from mathematics, statistics, and numerical analysis. In this technique the dataset is partitioned into two or more factions (clusters) of similar records. The clustering algorithms aim at grouping records keeping in mind the ultimate objective of maximizing a similarity metric between the members of the cluster. In most cases, closeness is the similarity metric and the aim is to maximize the cumulative closeness between data records in a cluster. The researchers then explore the properties of the members of the generated clusters. Association Models Association rule(X) Y is defined over a set of transactions T where X and Y are sets of items. In a Clinical setting, the set T can be patient's clinical records and items may be symptoms, measurements, observations, or diagnosis corresponding to the patients clinical records. Given S as a set of items, support(S) is defined as the number of transactions in T that contain all members of the set S. The confidence of a rule (X) Y is defined as $\text{support}(X(Y))/\text{support}(X)$, and the support of this rule is $\text{support}(X(Y))$. The discovered association rules show hidden patterns in the mined dataset. For example, the rule: ({People who are alcoholic})/ {People needing dialysis} with a high confidence signifies that the number of people requiring dialysis is high among people who are alcoholic. Healthcare professionals store significant amounts of patient data that could be used to extract useful knowledge. Researchers have been investigating the use of statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of heart disease. Statistical analysis has identified the risk factors associated with heart disease to be age, blood pressure, smoking habit [1], total cholesterol [2], diabetes [3], hypertension, family history of heart disease [4], obesity, and lack of physical activity [5]. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Researchers have been applying different data mining techniques such as decision tree, naïve bayes, neural network, bagging, kernel density, and support vector machine over different heart disease datasets to help health care professionals in the diagnosis of heart disease [4]-[5]. In[3] showed correct classified accuracy of approximately 77 % with logistic regression. The another model R-C4.5 is applied which is based on C4.5 shows better results, rules created by R-C4.5 can give health care experts clear and useful explanations [4].

III.METHODOLOGY USED

Feature Selection (FS) a preprocessing technique is used to identify the significant attributes, which play a dominant role in the task of classification. This leads to the dimensionality reduction. By applying different approaches features can be reduced. The reduced feature set improves the accuracy of the classification task in comparison of applying the classification task on the original data set. The overall procedure includes the following steps as shown in figure 1.1. Preprocessing of data which is in any format. 2. Selection of attributes using feature selection for dimensionality reduction 3. Dataset with reduced set of attributes given as input to the classifier. 4. Allocation of class.

Let S be the system:

$S = \{A, B, C\}$

Where



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

A=Set of input attributes for the medical dataset

B= Approach of feature selection

C= Set of selected attributes or set of removed attributes

B indicates the feature selection approach which may be either brute force approach, existing feature selection techniques or novel approach using correlated and associated rules.

Association is one of the most vital approach of data mining that is used to find out the frequent patterns, interesting relationships among a set of data items in the data repository. It is also known as market basket analysis due to its capability of discovering the association among purchased item or unknown patterns of sales of customers in a transaction database. For example if a customer is buying a computer then the chance of buying antivirus software is high. This information helps the storekeeper to further enhance their sales. Association also has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms. Ji et al., used association in order to discover infrequent casual relationships in Electronic health databases [8]. Healthcare organization widely used Association approach for discovering relationships between various diseases and drugs. It is also used for detecting fraud and abuse in health insurance. Association is also used with classification techniques to enhance the analysis capability of Data Mining. Soni et al., used an integrated approach of association and classification for analyzing health care data. This integrated approach is useful for discovering rules in the database and then using these rules an efficient classifier is constructed. This study performed experiment on the data of heart patients and also generate rules using weighted associative classifier. Bakar et al., also construct a predictive model using various rule based classifier for dengue occurrence. In this research work authors combine rough set, naïve bays, decision tree and associative classifier to build a predictive model for enhancing the early detection of dengue occurrence [9]. Doctor's prescriptions and treatment materials are produced large amount of data. Utah Bureau of Medicaid Fraud used this data to discover hidden and useful information in order to detect fraud. This approach is also helpful for identifying the improper prescriptions, irregular or fake patterns in medical claims made by physicians, patients, hospitals etc. in this figure2 and figure3 using classification rules.

Classification vs. association rules

- **Classification rule:**
predicts value of a given attribute (the classification of an example)

```
If outlook = sunny and humidity = high  
then play = no
```
- **Association rule:**
predicts value of arbitrary attribute (or combination)

```
If temperature = cool then humidity = normal  
If humidity = normal and windy = false  
then play = yes  
If outlook = sunny and play = no  
then humidity = high  
If windy = false and play = no  
then outlook = sunny and humidity = high
```

54

Figure 2: classification used association rules mining

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Example : Heart diseases Dataset

ATTRIBUTES		ID	age	Gender	Chest pain	Blood pressure	diagnosis	CLASS VALUES
INSTANCES		1	63	male	typ_angina	High	No	
		2	67	male	asympt	very_high	Yes	
		3	67	male	asympt	high	Yes	
		4	37	male	non_anginal	high	No	
		5	41	female	atyp_angina	high	No	
		6	56	male	atyp_angina	high	No	
		7	62	female	asympt	high	Yes	
		8	57	female	asympt	high	No	
		9	63	male	asympt	high	Yes	
		10	53	male	asympt	high	Yes	
		11	57	male	asympt	high	No	
		12	56	female	atyp_angina	high	No	
		13	56	male	non_anginal	high	Yes	
		14	44	male	atyp_angina	high	No	

Figure 3: Example of heart diseases dataset

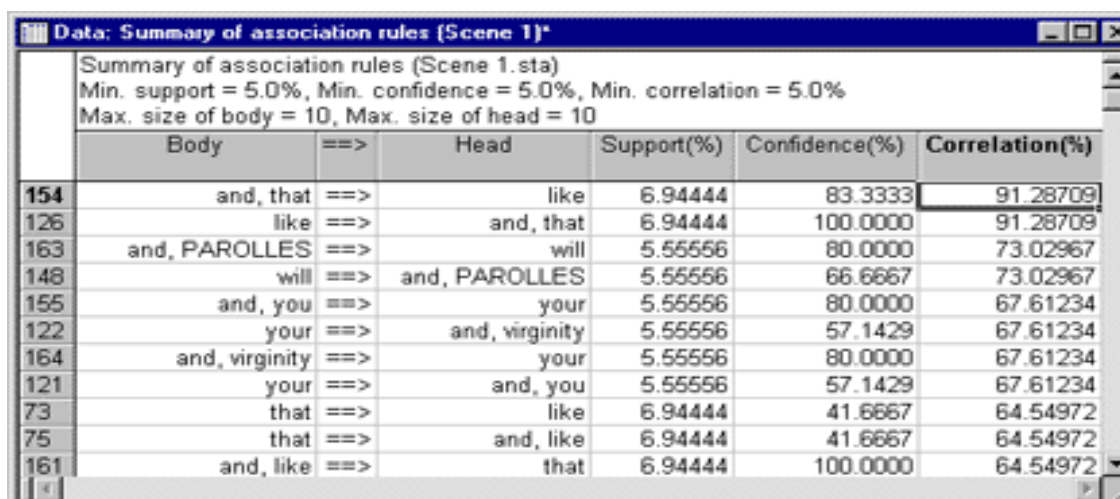
Attribute name	Medical meaning	Neg	Constraints		
			itemFilter	group	ac
AGE	Patient age	0	0	0	1
LM	Left Main	1	1	0	2
LAD	Left Anter Desc	1	1	0	2
LCX	Left CircumfleX	1	1	0	2
RCA	Right Coronary	1	1	0	2
AL	Antero-Lateral	0	1	1	1
AS	Antero-Septal	0	1	1	1
SA	Septo-Anterior	0	1	1	1
SI	Septo-Inferior	0	1	1	1
IS	Infero-Septal	0	1	1	1
IL	Infero-Lateral	0	1	1	1
LI	Latero-Inferior	0	1	1	1
LA	Latero-Anterior	0	1	1	1
AP	Apical	0	1	1	1
SEX	Gender	0	0	0	1

Figure 4: Example of data mining in medical data set attribute.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016



	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
154	and, that	==>	like	6.94444	83.3333	91.28709
126	like	==>	and, that	6.94444	100.0000	91.28709
163	and, PAROLLES	==>	will	5.55556	80.0000	73.02967
148	will	==>	and, PAROLLES	5.55556	66.6667	73.02967
155	and, you	==>	your	5.55556	80.0000	67.61234
122	your	==>	and, virginity	5.55556	57.1429	67.61234
164	and, virginity	==>	your	5.55556	80.0000	67.61234
121	your	==>	and, you	5.55556	57.1429	67.61234
73	that	==>	like	6.94444	41.6667	64.54972
75	that	==>	and, like	6.94444	41.6667	64.54972
161	and, like	==>	that	6.94444	100.0000	64.54972

Figure 5: Attribute value accuracy for using Association rules.

IV. EXPERIMENTAL RESULT

Association rules represent a promising technique to improve heart disease prediction. Unfortunately, when association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. In [10], four constraints were proposed to reduce the number of rules: item filtering, attribute grouping, maximum item set size, and antecedent/consequent rule filtering. When association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. A more important issue is that, in general, association rules are mined on the entire data set without validation on an independent sample. To solve these limitations, the author has introduced an algorithm that uses search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. Instead of using only Support and confidence, one more parameter i.e. lift have been used as the metrics to evaluate the medical significance and reliability of association rules. Medical doctors use sensitivity and specificity as two basic statistics to validate results. Sensitivity is defined as the probability of correctly identifying sick patients, where as specificity is defined as the probability of correctly identifying healthy individuals. Lift was used together with confidence to understand sensitivity and specificity.

In this figure 6: shows the accuracy of the algorithm obtained from experiment and then (fig-7) good graph model.

Data mining techniques	Accuracy
Naive bays	89.0%
Decision tree	91.7%
Classification And Association rules	96.9%

Figure: 6 Algorithm based accuracy



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

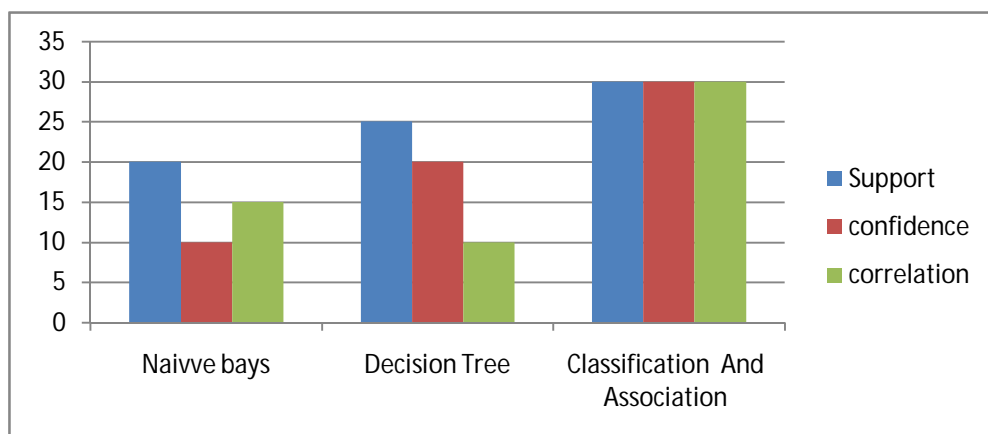


Figure: 7 Graph of accuracy

In this paper attributes the performance of classification can be improved and also cost of classification may gets reduced. In this work the analysis of two different approach for feature selection is done specially for medical datasets. This work also shows a novel approach for feature selection using correlation and by generating association rules. We conclude that feature selection really helpful for dimensionality reduction and also for building cost effective model for disease prediction. Our further work is to implementation of novel approach of using association rule mining and in future to develop a Cost effective Clinical Decision Support System for more accurate difficult cancer treatment.

REFERENCES

1. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, (1996) "From Data Mining to Knowledge discovery in Databases ", AI Magazine Volume 17 Number 3 (© AAAI)
2. Jiawei Han, Micheline Kamber, (2006)"Data Mining: concepts and Techniques", ELSEVIER.
3. John Shafer, Rakesh Agarwal, and Manish Mehta, (1996)"SPRINT: A scalable parallel Classifier for data mining", In Proc. Of the VLDB Conference, Bombay, India.
4. David Hanauer, "Mining clinical electronic data for research and patient care: Challenges and solutions", Clinical Assistant Professor, University of Michigan, USA, September 2007.
5. R. Agrawal et al., "Fast discovery of association rules, in Advances in knowledge discovery and data mining", MIT Press, pp. 307-328, 1996.
6. Medline Resources <http://www.nlm.nih.gov/bsd/pmresources.html>
7. Lalayants et.al, "Clinical data-mining: Learning from practice in international settings", International Social Work, doi: 0020872811435370, March 27, 2012
8. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", VLDB, Chile, ISBN 1-55860-153-8, (1994) September 12-15, pp. 487-99.
9. J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, "Mining Infrequent Causal Associations in Electronic Health Databases", 11th IEEE International Conference on Data Mining Workshops, (2011).
10. Carloz Ordonez, Association Rule Discovery with Train and Test approach for heart disease prediction, IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006, pp 334-343