Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining

Chetana Yadav
M.Tech Scholar
Department of IT
RKDFIST, Bhopal

Shrikant Lade
Asst. Professor
Department of IT
RKDFIST, Bhopal

Manish K Suman
Asst. Professor
Department of IT
RKDFIST, Bhopal

ABSTRACT

In this paper, we present an improved association rule mining of data mining for the detection of Coronary Artery Disease (CAD). The whole concept behind the research is carried out by using a heart disease database, which is collected from different locations and from different patients. It is actually collected from 303 random visitors to Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre. The mechanism proposed in this article uses the same heart disease database as input and apply the improved association rule mining method to identify the decision rules with the correctness and robustness. This paper shows conclusions of our proposed work and results.

1. INTRODUCTION

The heart disease is a major cause of mortality and death-rate in modern society. Medical assessment is very important but labyrinthine task that should be performed efficiently and accurately. Therefore, in this era of computing and intelligence. It is an easy yet complicated task to estimate the probability of disease on the basis of data and fact provided to the system. The morality rates from diseases are much greater than those of accidents and natural disasters. The World Health Organization estimates that 17 million deaths worldwide each year occur due to cardiovascular diseases. A major type of such diseases is coronary artery disease (CAD), which is reported to account for 7 million deaths over the world per annum.

In data mining, association analysis is a method for discovering interesting relationships or patterns between variables [4]. The Apriori and FP-growth association algorithms are the two most efficient association algorithms for a data set the size of ours [4,5]. Their efficiency is based on generating frequent item sets and rules based on support level and confidence measures.

The association rule mining algorithms are typically used to identify the patterns that occur in original form ubiquitously the database. In any database that contains many minor variations in data values, potentially vital finding may be disregarded as a result.

For the definition of data mining, now there are two perspectives: some people think that data mining is knowledge discovery in databases; another group of people who think data mining is the process of knowledge discovery in databases, a processing step. According to Fayyad's definition: Data mining is found from the database potentially useful, novel, understandable model of high-level processing. As can be seen from the definition, data mining is knowledge discovery in databases process the core of a step. Knowledge discovery in databases, including the following four main steps:

•The original data selection and cleaning: According to data mining purposes, determine the need for data mining, And raw data noise and missing values for processing.

• Data transformation: In order to meet the needs of data mining algorithms, data representation or data layout transformations, such as continuous numerical attribute discretization, the level of the layout of the data is transformed into vertical layout.

• Data Mining: Use of specific algorithms to extract knowledge from data, knowledge representation, including concepts, rules and patterns.

•Model assessment: Evaluation of the importance of knowledge, confidence and fun degree, on the model to explain the reasonableness of assessment models and availability.

In 2008, M.Hussein et al. [6] proposed a scheme, MHS, also for privacy-preserving association rules mining on horrizontally distributed databases. MHS gives accurate mining results and has better privacy and performance than KCS [7] and V.E.Castro et al.'s scheme [8]. Particularly, MHS satisfies semi-honest model; only if Initiator and Combiner colluding with each other can reveal the secret data in malicious mode; and it takes only 3 steps in each phase.

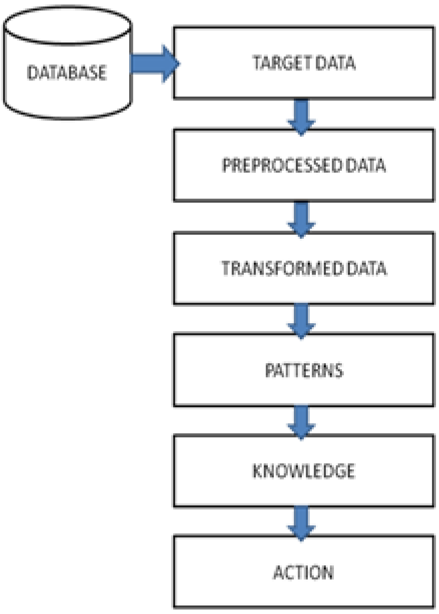


Figure 1: Data Mining Process

1.1 Used Medical Dataset

The Z-Alizadeh Sani dataset is collected from 303 random visitors to Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center and contains 54 features [1]. The features along with their valid ranges are given in Tables 1, 2

Table 1: Demographic Features

Demographic Features	Range
Age	30-86
Weight	48-120
Sex	Male, Female
Body Mass Index (Kg/m^2)	18-41
Diabetes Mellitus	Yes/No
Hypertension	Yes/No
Current Smoker	Yes/No
Ex Smoker	Yes/No
Family History	Yes/No
Obesity	Yes, if MBI>25 Otherwise No
Chronic Renal Failure	Yes/No
Cerebrovascular Accident	Yes/No
Airway Disease	Yes/No
Thyroid Disease	Yes/No
Congestive Heart Failure	Yes/No
Dyslipidemia	Yes/No

Table 2: Laboratory Features

Laboratory Features	Range
FBS (Fasting Blood Sugar)	62-400
Cr (Creatine)	0.5-2.2
TG (Triglyceride):	37-1050
LDL (Low-Density Lipoprotein)	18-232
HDL (High-Density Lipoprotein)	**15-111**
BUN (Blood Urea Nitrogen)	6-52
ESR (Erythrocyte Sedimentation Rate)	1-90
Hb (Hemoglobin)	8.9-17.6
K (Potassium)	3.0-6.6
Na (Sodium)	128-156
WBC (White Blood Cell)	3700-18000
Lymph (Lymphocyte)	7-60
Neut (Neutrophil)	32-89
PLT (Platelet)	25-742
EF (Ejection Fraction)	15-60
Region with RWMA (Regional Wall Motion Abnormality)	0/1/2/3/4
VHD (Valvular Heart Disease)	Normal/Mild /Moderate/Severe

2. LITERATURE SURVEY

This section of the paper represents the existing work and research in the field of data mining. Association rules are conditional statements that help to reveal relationships between distinct data items in a relational database or other data storage. An association rule has two parts, antecedent and consequent. First part is a set of items found in the transactional data. Second part is an itemset that is present in combination with the antecedents. An association rule is an implication of the type $P \rightarrow Q$, where P and Q are itemsets, which are disjoint in nature. i.e., $P \cap Q = \emptyset$. Generally, the power of the association rule is determined by its objective parameters such as support and confidence. Support determines the percentage of transaction containing both antecedent and consequent in the dataset. Confidence determines the ratio of the number of occurrence of both antecedent and consequent to the number of occurrence of antecedent in the data set.

In the study [1], the MetaCost algorithm, which is a cost-sensitive algorithm, was used. First, from a total of 54 features, 34 were selected using feature selection algorithm. Then three created features were added to the dataset. The C4.5, KNN, Naïve Bayes, SVM and SMO algorithms were thereafter used in MetaCost. The accuracy of the SMO algorithm was better than that of the other Algorithms. The accuracy of the KNN is not as well as the other algorithms since the number of patients who have CAD is about 2.5 times more than the number of normal ones and also comparison of Euclidean distance between patients cannot accurately discriminate them so this algorithm is more likely to diagnose patients as CAD. In order to study the cost sensitive algorithms, first the cost matrix was set with no difference between the two classes. Next, taking two times and third times the cost for the wrong CAD diagnosis, it was seen that third case leaded to the best sensitivity. In addition, the feature creation method was investigated. This method increased the accuracy of some of the classification algorithms, substantially.

In the research [2] presents a particle swarm optimization (PSO)-based fuzzy expert system for the diagnosis of coronary artery disease (CAD). The designed system is based on the Cleveland and Hungarian Heart Disease datasets. Since the datasets consist of many input attributes, decision tree (DT) was used to unravel the attributes that contribute towards the diagnosis. The output of the DT was converted into crisp if–then rules and then transformed into fuzzy rule base. PSO was employed to tune the fuzzy membership functions (MFs). Having applied the optimized MFs, the generated fuzzy expert system has yielded 93.27% classification accuracy. The major advantage of this approach is the ability to interpret the decisions made from the created fuzzy expert system, when compared with other approaches.

In the research [3], Coronary artery disease (CAD) affects millions of people all over the world including a major portion in India every year. Although much progress has been done in medical science, but the early detection of this disease is still a challenge for prevention. The objective of this paper is to describe developing of a screening expert system that will help to detect CAD at an early stage. Rules were formulated from the doctors and fuzzy expert system approach was taken to cope with uncertainty present in medical domain. This work describes the risk factors responsible for CAD, knowledge acquisition and knowledge representation techniques, method of rule organisation, fuzzification of clinical parameters and defuzzification of fuzzy output to crisp value. The system implementation is done using object oriented analysis and design. The proposed methodology is developed to assist the medical practitioners in predicting the patient's risk status of CAD from rules provided by medical experts. The present paper focuses on rule organisation using the concept of modules, meta-rule base, rule address storage in tree representation and rule consistency checking for efficient search of large number of rules in rule base. The developed system leads to 95.85% sensitivity and 83.33% specificity in CAD risk computation.

In research [9] the authors presented a predictive model for the Ischemic Heart Disease (IHD); they applied Back-propagation neural network (BPNN), the Bayesian neural network (BNN), the probabilistic neural network (PNN) and the support vector machine (SVM) to develop classification models for identifying IHD patients on a data obtained from measurements of cardiac magnetic field at 36 locations (6 × 6 matrices) above the torso.

3. PROPOSED METHOD

The work presented in this paper is the application of data mining for discovering hidden unknown knowledge from a medical dataset. The Medical data is temporal in nature and therefore traditional data mining techniques are not appropriate. This dataset contains medical records of coronary artery disease patients. The structure of these medical records is chain of observations taken at different times. In each observation, a set of clinical parameter are saved. The aim of this paper is mining relational rules from this set of medical data that can be used in early prediction and of risk in the patients. In the first part of this study a pre-processing technique is used to produce primary structure of medical records. On which different data mining techniques is applied.

In the proposed implementation [10], typical steps of the data mining process are used which is used as hypothesis generator. While in this paper, the effectiveness of an association rule-mining algorithm on the dataset is investigated.

Pseudo code of Improved Apriori using Transaction Reduction Method:

```
min-sup-count = min-sup * |D|
$L_1$-candidates = find all one itemsets (D)
//ScanDandproduce$L_1$-candidates
$L_1$={<$X_1$,TR-set($X_1$)>L∈1-candidates|sup-
                              count≥min-sup-count}
for (k=2; $L_{k-1}$≠ϕ; k++) {
  do {
    for each k-itemset($X_i$,TR-set($X_i$) ∈ $L_{k-1}$ do
    for each k-itemset($X_j$,TR-set($X_j$)) ∈ $L_{k-1}$ do
    if($X_i$[1] = $X_j$[1]) ∧ ($X_i$[2] = $X_j$[2]) ∧…
       ........∧ ($X_i$[k-2] = $X_j$[k-2]) then  {
      $L_k$-candidates.$X_k$ = $X_i$ * $X_j$;
      $L_k$-candidates.TR-set($X_k$)=TR-set($X_i$)∩TR-set($X_j$)
      }}
for each k-itemset <$X_k$,TR($X_k$)> ∈ $L_k$-candidates do
sup-count = |TR-set|
$L_k$ = {<$X_k$,TR-set($X_k$)> ∈ $L_k$-candidates |
                               sup-count ≥ min-sup-count}
}
set-count = $L_k$.itemcount            //updataset-count
return  L = $\cup_k L_k$;
```

In the proposed method, the dataset is properly design according to the compatibility with the data mining algorithms, after observing the dataset, weight must be assigned to the feature and then the features are selected for the task of classification according to their weights and finally the association rules will be generated.

we propose the improvement to the traditional Apriori algorithm by applying following conditions:
(1) Compute minimum sup-count by min-sup*|D|.

(2) Scan the database once to produce L_1-candidates, simultaneously construct TR-set(X_1) for each item. After scanning, compute sup-count for each item and find the set of frequent items L_1.

(3) Produces L_2-candidates from L_1*L_1.Scanning L_1, we can find TR-set(X_2) and sup-count of each item set, deletes the

patterns whose frequencies don’t satisfy the min support count, and find L_2.

(4) In order to produce $L_k(k\geq3)$, join itemsets which satisfy the join rule. Scanning Lk-1, we can find TR-set(X_k) and compute sup-count of each itemset, then deletes the patterns whose frequencies do not satisfy the minimum support count, and find L_k.

In this way, the interesting association rule can be accessed more effectively by the improved Apriori algorithm.

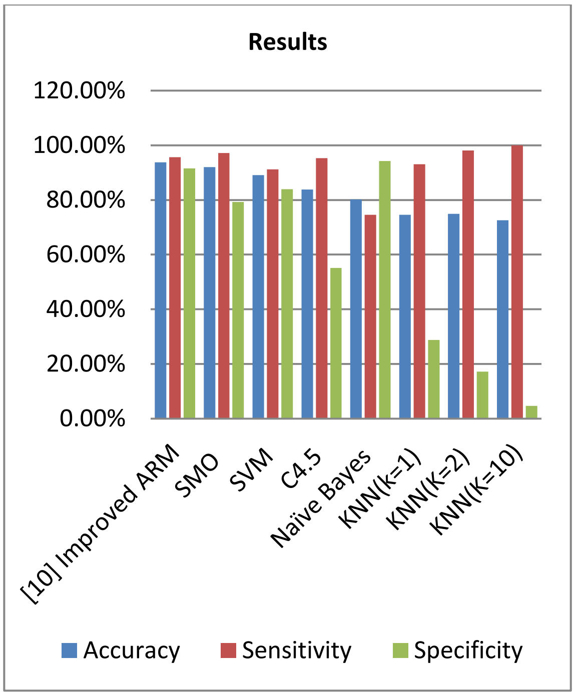
4. EXPERIMENTAL RESULTS

The proposed approach of data mining is carried out in C#.NET. Microsoft .NET is software that connects information, people, systems and devices. The .NET Framework is the programming model of the .NET environment for building applications. It manages much of the plumbing, enabling developers to focus on writing the business logic code for their applications. The .NET Framework includes the common language runtime and class libraries.

In this section, we evaluate the performance of our proposed method under heart disease dataset. The experimental results demonstrate that the proposed algorithm perform well in terms of the accuracy, sensitivity and specificity.

Algorithm	Accuracy	Sensitivity	Specificity
Improved ARM [10]	93.75%	95.65%	91.53%
SMO	92.09%	97.22%	79.31%
SVM	89.11%	91.20%	83.91%
C4.5	83.85%	95.37%	55.17%
Naïve Bayes	80.15%	74.54%	94.25%
KNN(k=1)	74.61%	93.06%	28.74%
KNN(K=2)	74.94%	98.15%	17.24%
KNN(K=10)	72.62%	100%	4.6%

Following is the graphical representation, that shows the performance of various data mining algorithms; The accuracy, sensitivity and specificity is determined for each algorithms based on the values of various ROC parameters and the formulas.



Apriori algorithm is improved based on the properties of reduction in transactions. The typical Apriori algorithm has performance bottleneck in the massive data processing so that we need to optimize the algorithm with variety of methods. The improved algorithm we proposed in this paper not only optimizes the algorithm of reducing the size of the candidate set of k-itemsets, Ck, but also reduce the I/O spending by cutting down transaction records in the database. The performance of Apriori algorithm is optimized so that we can mine association information from massive data faster and better.

5. CONCLUSION & FUTURE WORK

The application of Association Rule Data Mining for identifying the hidden knowledge from medical dataset is proposed in this paper. The medical data is temporal in nature and hence traditional data mining techniques are not suitable. This dataset used for the work proposed in this paper contains medical records of Coronary Artery Disease. The generated association rule patterns from this dataset were presented to medical experts in the field.

After careful evaluation on correctness and applicability of the results by these experts positive feedbacks were received.

In future, we will emphasize on diagnosis of Breast Adenocarcinoma also further tests can be carried out on the same approach for deeper penetration in the diagnosis. This work can further pave way for detection of cancer and other diseases and consequently impact the way such diseases are diagnosed.

6. REFERENCES

[1] Roohallah Alizadehsani, Mohammad Javad Hosseini, Zahra Alizadeh Sani, Asma Ghandeharioun, Reihane Boghrati, “Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms”, IEEE 12th International Conference on Data Mining Workshop, 2012, pp.9-16.

[2] S. Muthukaruppan, M.J. Er, “A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease”, Expert Systems with

Applications, Volume 39, Issue 14, 15 October 2012, Pages 11657–11665

[3] Debabrata Pal, K.M. Mandana, Sarbajit Pal, Debranjan Sarkar, Chandan Chakraborty, “Fuzzy expert system approach for coronary artery disease screening using clinical parameters”, Knowledge-Based Systems, Volume 36, December 2012, Pages 162–174

[4] P.N. Tan, M. Steinbach, V. Kumar. Introduction to Data Mining. 2006. Pearson Addison-Wesley, Boston, MA

[5] R. Hu. Medical Data Mining Based on Association Rules. Computer and Information Science.Vol. 3, 2010, No. 4

[6] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail: Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base. Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607 -- 616 (2008)

[7] Murat Kantarcioglu and Chris Clifton: Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. IEEE transactions on knowledge and data engineering - Volume 16 Issue 9, September 2004 (2004)

[8] Vladimir Estivill-Castro, Ahmed HajYasien: Fast Private Association Rule Mining by A Protocol for Securely Sharing Distributed Data. Intelligence and Security Informatics, IEEE, pp. 324 -- 330 (2007) .

[9] Kangwanariyakul, Y., Chanin, N., Tanawut, T., Thanakorn, N. “Data Mining of Magnetocardiograms for Prediction of Ischemic Heart Disease”, EXCLI Journal. 33(9) 2010, Pp.:82-95.

[10] Chetna yadav, A Survey on Data Mining Techniques for the diagnosis of Coronary Artery Disease, 2013 IJARCSSE Volume 3 Issue 10.