Journal of Theoretical and Applied Information Technology
20th September 2015. Vol.79. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ANALYSIS AND ACCURACY PREDICTION OF HEART DISEASE WITH ASSOCIATION RULE AND OTHER DATA MINING TECHNIQUES

[1] SHABANA ASMI P, [2]Dr. S. JUSTIN SAMUEL
[1]Asst. Prof., Department of Computer Applications, Faculty of Computing, SathyabamaUniversity
[2]Professor, Department of Information Technology, Faculty of Computing, SathyabamaUniversity
E-mail: [1]shabhanaasmi@yahoo.com, [2]drsjustin@gmail.com

ABSTRACT

Disease is a term for defining a large amount of healthcare conditions that affects part or all of an organism. Knowledge of various disease symptoms and signs favorable for disease development is necessary to optimize a disease forecaster. Different data mining techniques such as Naive Bayes, Decision Tree, Linear Regression and Association Rule are used to predict the heart disease. Data mining techniques in all disease diagnosis applied over all disease treatment dataset investigate if hybrid data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis. In this paper, our work is to more accurately predict the presence of heart disease with added attributes of the disease and using association rules. Final results show that association rule implemented on dataset produces better accuracy.

Keywords: *Artificial Neural Network (ANN), Cleveland Heart Disease Data Set (CHDD), k-Nearest Neighbor (KNN), Heart Disease Prediction and Diagnosis, Knowledge Discovery in Database (KDD)*

1. INTRODUCTION

Information mining is the examination of far reaching datasets to think covered and in advance dark samples, associations and learning that are tricky to recognize with conventional strategies. Information mining is quickly developing effective in an extensive variety of utilizations like, examination of natural mixes, budgetary anticipating, medicinal services and climate determining. Information mining in health awareness is in the rising field of significance in giving visualization and the broader cognizance of remedial data. The Economical Commission of Asia reported that in twenty percent of Asian countries, most deaths are due to non-transferable ailments such as cardio afflictions, growths, diabetes and incessant respiratory maladies.

The Australian Statistics Bureau stated that heart and the circulatory system sicknesses are one of the first main reason for death in Australia, bringing on nearly 33.7% passing's. Measurements of South Africa stated the report that heart and circulatory system ailments are the 3rd driving reason for passing's in Africa.

Adding to a device to be installed in the clinics, administration framework to help and offer guidance to the social insurance experts in identifying and providing the suitable treatment for coronary illness in people are critical. A few information mining procedures [1] like Decision Tree, Naïve Bayes, Neural system, k-Nearest Neighbor (KNN) [2], Piece thickness, Automatic character gatherings, Stowing calculation are utilized as a part of the analysis of these coronary illnesses and they are used in the vector machine displaying deeper levels of correctness [3].

1.1 Existing System

Clinical choices are frequently made in view of specialists' hunch and strife instead of the information independent information gushed up in the database. There are numerous ways that a medicinal misdiagnosis can display it. Whether a specialist is at short coming, or doctor's facility staff, a misdiagnosis of a genuine disease can have exceptionally compelling and destructive impacts. The National Patient Safety Foundation refers to that 42% of therapeutic patients feel they have had encountered a restorative lapse or missed the conclusion. Tolerant security is now and then carelessly given the secondary lounge for different concerns, for example, the expense of medical tests, medications, and operations. Restorative Misdiagnoses are a genuine danger to our human services calling. On the off chance that they proceed with, then individuals will apprehension

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

setting off to the healing center for treatment.

2. RELATED WORK

Factual Pattern Recognition: A Review

The objective of the example is directed or undirected grouping. Amidst the separate structures where the design distinguish has custom formed, the measurement methodology has been most seriously mulled over and utilized as a part of practice. All the more as of late, neural system methods and strategies transported in from measurable learning have been getting the consideration. Outline of a distinguished framework needed watchful observation regarding the issues: meaning of the example sensing environment, classes, design representation, choice, characteristic extraction and bunch examination, plan and learning, the preparing tests and assessment. Disregarding very nearly 50 years of inventive work in this area, the normal issue of seeing the complicated examples with the discretionary introduction, scale, area stays unsolved. Newer applications, for example, information mining, web seeking, recovering the interactive media information's, face detection and the cursive penmanship are the examples for distinguished strategies. The goal of this paper is to condense and think about a percentage of the remarkable routines utilized as a part of different phases and to recognize research themes and applications which are at the bleeding edge of this energizing and testing field.

Information Mining in Healthcare: Current Applications and Issues

The fruitful utilization of information mining in exceptionally obvious in the fields like e-business, advertising and retail have prompted the ubiquity of its utilization in learning, revelation in databases, Knowledge Discovery in Database (KDD) is used at different businesses and segments. Among these simple information mining is applied in the fields of drug and general wellbeing. It also discusses critical issues and challenges associated with data mining and health care in general [4]. The research found a growing number of data mining applications, including analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims.

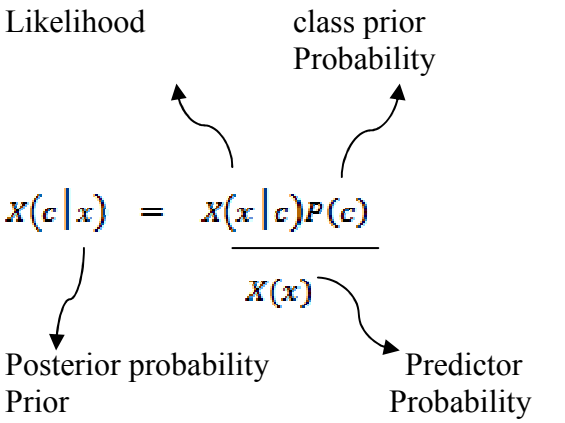
2.1 Proposed System

In the proposed System, we are applying data mining techniques (Hybrid) in identifying suitable treatments for all disease patients. Apply multiple data mining techniques to all disease diagnosis benchmark datasets to establish baseline accuracy for each single data mining technique in the diagnosis of all disease patients. The same multiple data mining techniques used in all disease diagnosis to all disease treatment dataset to investigate if single data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis. In this paper, we proposed a data mining technique called Association Rule, which leads to succeed with less time consuming and achieved by high performance and accuracy.

3. ALGORITHMS USED

3.1 Naive Bayes

Bayes theorem provides a way of calculating the posterior probability, X (c/x), from X (c), X (x), and X (x/c). [5] Naive Bayes classifier assumes, that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



$X(c|x) = X(x_1|c) \times X(x_2|c) \times \ldots\ldots \times X(X_n|c) \times X(c)$

X(c | x) is the posterior probability of class (target) given Predictor (attribute)

X(c) is the prior probability of class.

X(x | c) is the likelihood which is the probability of predictor given class.

X(x) is the prior probability of predictor.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2 Decision Trees

The Decision Tree calculation [6], in the same way as Naive Bayes, is taking into account contingent probabilities.

Sample Algorithm:
Top-down Algorithmic Framework for Decision Trees Induction

Terms	Explanation
W, H, Y	Training set, Input Feature set, Target feature
SM	Splitting Metric
T	Threshold

Procedure Tree Growing (W, H, Y)
If (SM>T) then
Label t with f (A)
In (; each outcome view of f (A;)
{
Set Subtree$_i$=TreeGrowing ($\sigma_{f(A)=v_i}$ W, H, Y)
Root node = t_T
label as v_i
}
Else
{
Root node =T as a leaf
Label S
}
End if
Return T

Terms	Explanation
W	Training set
H	Target feature
O	Pruned tree

Procedure Tree Pruning (W, H, O)

Do
Select a node t in T
If t≠ϕ then T=pruned (T, t)
Until t≠ϕ
Return T

Table.1: List of attributes for lung cancer

Id	Attribute Name	Description
1	Age	Age in Year
2	Sex	1: Male; 0: Female
3	Hoarseness	1. Yes, 2. No
4	Radon	1. Yes, 2. No
5	Asbestos	1. Yes, 2. No
6	Tobacco Smoking	1. Yes, 2. No
7	Marijuana	1. Yes, 2. No
8	Recurring Inflammation	1. Chronic 2. Acute
9	Family history	1. Mother, sister, or daughter had Stomach cancer 2. None
10	Chest pain type	1. Angina 2. Asthma and Other Respiratory Conditions 3. Coronary artery spasm 4. Atypical chest pain

3.3 Neural Networks

Neural network (NN) is a mathematical model which consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. [7] A neural network has to be configured such that the application of a set of inputs produces the desired set of outputs. This is an extremely basic model and comprises of a solitary trainable' neuron. Non-linear predictive models or Artificial Neural Network (ANN) models that learn through training set and resemble biological neural networks in structure. Inputs are exhibited to the neuron and every detail has a coveted yield. In most cases, neural networks perform as well or better than the traditional statistical techniques to which they compared [8]. Thus, neural networks are becoming

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

very popular with data mining practitioners, particularly in medical research. Set of attributes mainly used for lung cancer is listed in table 1 [9].

3.4 Association Rule

Given a set of transaction T, the goal of association rule mining is to find all rules having support ≥ minsup threshold, confidence ≥ minconf threshold.

Computational Complexity

Given d unique items, Total number of item sets=2^d

Total number of possible association rules

$$R = \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right]$$

3.5 Regression Rules

Regression Rule takes a numerical dataset and develops a mathematical formula,

$$m = a + bn$$

Here m is the dependant variable and n is the independent variable that fits the information. A major limitation in this technique is that, it only works well with continuous quantitative data (like weight, speed or age). If data are categorical, where order is not significant (like color, name or Gender) then it is better off choosing another technique.

4. COMPARISON WITH EXISTING DATA MINING TECHNIQUES

As of late, specialists began utilizing mixture information mining methods as a part of the analysis of coronary illness. Polat et al., Utilized fluffy counterfeit safe distinguishment framework and k-closest neighbor in the discovery of coronary illness utilizing Cleveland Heart Disease Data Set (CHDD). The model demonstrated precision of 87% in the area of the coronary illness patients (Table 2). [10]-[11] Utilized neural system troupes as a part of the judgment of coronary illness utilizing the CHDD indicating exactness of 89.01% precision (Table 2). This model proposes the neural system prepared utilizing the hereditary calculation demonstrating that the hybridization is steady. This exploration did not compute model precision.

Table 2: A Sample Of Data Digging Techniques Used On The Cleveland Heart Disease Dataset

Type	Author/ Year	Technique	Accuracy
Single	Cheung	Decision Tree	81.11%
		Naive Bayes	81.48%
	Tu, et al., 2009	J4.8 Decision Tree	78.9%
		Bagging algorithm	81.41%
Hybrid	Polat et al., 2007	Fuzzy-AIRS–k-nearest Neighbor	87%
	Das, et al., 2009	Neural network ensembles	89.01%

Correlation of individual and crossover information collective methods in conclusion of coronary illness in CHDD shows distinctive correctnesses, with half breed strategies demonstrating preferred exactness over single procedures (Table 1). Hybridized information mining procedures are upgrading the precision of coronary illness determination. [12] The CHDD to exhibit that utilizing more modern information mining methods enhances the precision of coronary illness determination. Through a methodical examination of a few single information mining procedures, diverse information discretized levels, and diminished mistake pruning and the use of voting methods, we showed increments in exactness on all systems evaluated.

Table 3: Example Output Of Increased Accuracy On Bench Mark Dataset Using More Seasoned Data Digging Techniques

		Sensitivity	Specificity	Accuracy
Shouman et al., 2011	Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree	77.9%	85.2%	84.1%
Tu et al., 2009	J4.8 Decision Tree	72.01 %	84.48%	78.9%
	Bagging Algorithm	74.93 %	86.64%	81.49 %

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

While not flawless (clearly), these outcomes are great. The capacity to enhance the precision through the use of more advanced strategies is additionally empowering. The comparison uses more seasoned data digging technique and increased accuracy is determined using a decision tree algorithm.

Data Digging Techniques Used On The Cleveland Heart Disease Dataset

To determine the accuracy, Data mining technique is one of the common sectional approach. In this paper, we explore Bagging algorithm, where, bagging works as a method of increasing accuracy. The concept of bootstrapping and aggregation is based on Bagging. The method bootstrap samples the given training tuples uniformly with replacement. Bagging can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple.

5. PROPOSED METHODOLOGY

Naive Bayes, neural networks, decision tree, logistic regression, association rule is the data mining algorithms applied to the dataset [13,15]. Each and every algorithm is applied to the dataset and the result is produced. After implementing these algorithms into the dataset, a pattern for each dataset is created. Every algorithm will create each pattern for each database. Pattern formation is the process of creating the certain binary values according to the algorithm which is implemented. The input attributes are converted into some pattern by using data mining analysis algorithms. That pattern is matched to the pattern which is generated by the algorithm implementation in dataset by using mining process. After mining process it will produce some binary values for both dataset and input attributes. The binary values are generated for each algorithm implementation. The output shows the maximum probability percentage of heart disease and multi disease prediction system. By using this algorithm implementation, we can find accurate result of diagnosing the disease.

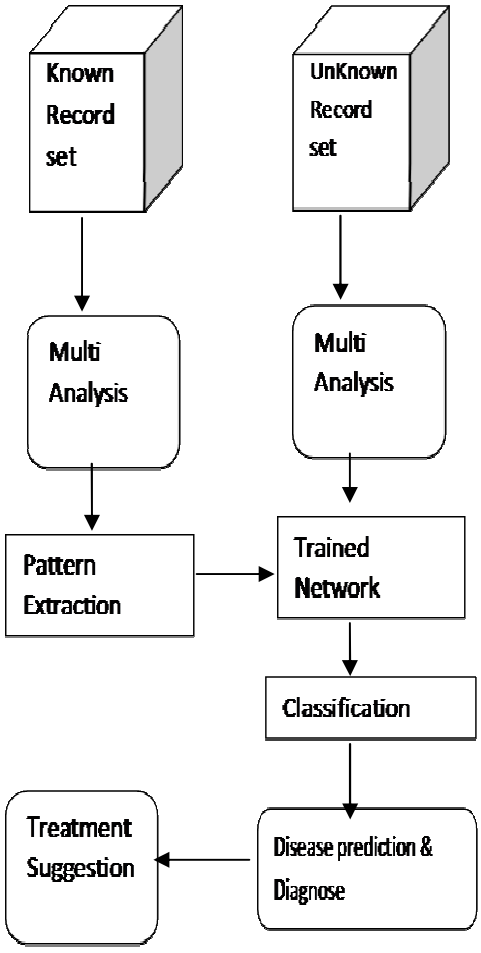


Figure 1: Proposed Research Model

6. EXPERIMENTAL RESULTS

The following results show the accuracy of different data mining algorithms using different support cases. [14] The accuracy of each method is plotted on a graph as below where pink bar represents accuracy for Instructed Heart Disease and resulted for proposed technique, called association rule. The other techniques, namely Naive Bayes, Decision tree, Neural networks, Regression Rule are suitable for predicting heart disease and finally it is compared with association rules. The efficiency and effectiveness of our proposed technique are validated by experiments on real life databases, we show that the association rule approach efficiently discovers accurate results.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Instructed Heart Disease Prediction System Using Data Mining Technique

	Naive Bayes	Decision Tree	Neural networks	Linear regression	Association Rule
Probability %	99.6	52.6	50.1	94.6	100
#Support Cases	296	264	204	293	297

Multi Disease Prediction System Using Data Mining Techniques

	Naive Bayes	Decision Tree	Neural networks	Linear regression	Association Rule
Probability %	94.5	75.1	87.4	99.6	100
#Support Cases	281	274	225	279	298

7. CONCLUSION AND FUTURE WORK

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, we analyzed dataset with attributes of Lung cancer to get more accurate results. Five data mining techniques were applied, namely Naive Bayes, Decision Tree, Neural Networks, Linear Regression and Association Rule. From results it has been seen that Association Rule provides accurate results as compare to Naive Bayes, Decision Tree, Neural Networks, Linear Regression. As a future work, it can be planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

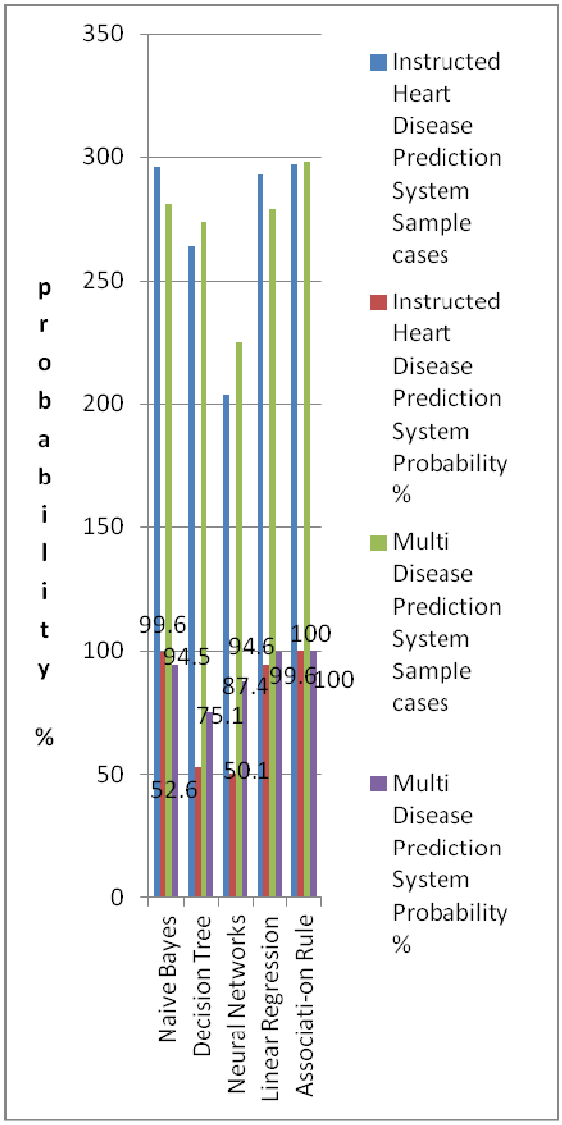


Figure 2: Graphical representation of accuracy for each method

REFRENCES:

[1] Han, j. and M. Kamber, Data Mining Concepts and Techniques. 2006: Morgan Kaufmann Publishers.

[2] Indira K, Christal Joy E, "Prevention of spammers and Promoters in Video Social Networks using SVM-KNN", International Journal of Engineering & Technology, Vol. 6, pp. 2024-2030,2014.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] Lee, I.-N., S.-C. Liao, and M. Embrechts, Datamining techniques applied to medical information. Med.

[4] Obenshain, M.K., Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 2004.

[5] http://www.saedsayad.com/naive_bayesian.htm

[6] www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

[7] Sandhya, J., et al.,Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques. International Journal of Engineering and Technology, 2010. Vol.2, No.4.

[8] Yashpal Singh, Alok Singh Chauhan Neural Networks In Data Mining, Journal of Theoretical and Applied Information Technology, 2005

[9] European Public Health Alliance. 2010 7-February-2011]; Available from: http://www.epha.org/a/2352

[10] World Health Organization. 2007 7-Febuary 2011]; Available from: http://www.who.int/mediacentre/factsheets/fs310.pdf.

[11] Thuraisingham, B., A Primer for Understanding and Applying Data Mining. IT Professional IEEE, 2000.

[12] Liao, S.-C. and I.-N. Lee, Appropriate medical data categorization for data mining classification techniques. MED. INFORM., 2002. Vol. 27, no. 1, 59–67, .

[13] Panzarasa, S., et al., Data mining techniques for analyzing stroke care processes. Proceedings of the 13th World Congress on Medical Informatics, 2010.

[14] Mai Shouman, Tim Turner Rob Stocker , Using Data Mining Techniques in Heart disease Diagnosis and Treatment ,School of Engineering and Information Technology University of New South Wales at the Australian Defence Force Academy .

[15] A. Viji Amutha Mary, S. Justin Samuel and D. Johnny Rajam, Automated Trinity Based Web Data Extraction for Simultaneous Comparison, Contemporary Engineering Sciences, Vol. 8, 2015, no. 11, 491 – 497.