Network and Complex Systems
ISSN 2224-610X (Paper) ISSN 2225-0603 (Online)
Vol.3, No.10, 2013

A Comparative Study of CN2 Rule and SVM Algorithm and Prediction of Heart Disease Datasets Using Clustering Algorithms

Ramaraj.M[*1], Dr.Antony Selvadoss Thanamani*[2]

[1]Research Scholar Department of Computer Science NGM College Pollachi India
ramaraj302@gmail.com.
[2]Associate Professor & Head Department of Computer Science NGM College Pollachi India
selvadoss@yahoo.com

Abstract

In this paper, we discuss diagnosis analysis and identification of heart disease using with data mining techniques. The heart disease is a major cause of morbidity and mortality in modern society; it is extremely important but complicated task that should be performed accurately and efficiently. It is an huge amount data of leads medical data to the need for powerful data analysis tools are availability on the data mining technique. They have long to been an concerned with applying for statistical and data mining tools and data mining techniques to improve data analysis on large datasets. In this paper, to proposed system are implemented to find out the heart disease through as to compared with the some data mining techniques are Decision tree, SOM, CN2 Rule and K-Means Clustering the data mining could help in the identification or the prediction of high or low risk of Heart Disease.

Keywords: Data Mining, Heart Disease, Decision tree, SOM, CN2 Rule and cluster.

1. Introduction

1.1. Overview of data mining

The data mining concern of database technologists was to find efficient method of storing, retrieving and falsify data, The main concern of the machine learning community was to develop this techniques for learning knowledge form data, Data mining was a marriage between technologies developed in the database and machine learning communities and the data mining can be considered to be an inter-disciplinary field involving from the databases to access in data to be a data mining concept as learning algorithms, database technology, statistics ,mathematics, clustering and visualization among others. It is existing real world data rather than data generated particularly for the learning tasks [1]. In data mining the data sets are large therefore efficiency and scalability of algorithms is important. A convention definition of KDD is given as follows:-data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data [2].These technology is to provides a user oriented approach to novel and hidden patterns in the data.

1.2. Heart Disease

Heart disease or heart disease is the class of diseases that involve the heart or blood vessels(arteries and veins).now today most countries face high and increasing rate of heart disease and it is become leading cause of debilitation and death in the Worldwide for men and women. The heart disease is affected by the men and women for the age of fifty to fifty-five years of the people to attack the heart disease.Harmonize to the World Health Organization report global atlas on heart disease prevention and control states that Heart Disease(CVDs) are the leading causes of death and disability in the world

- **A piece of information for heart diseases:**
- The number one cause of death globally as for CNDs;
- They more die on every year from this disease and any other cause[3]. And an estimated for 20% to 25% million people died from CVDs in 2011, representing 35% to 40% of all global death.

- In another diseases to estimate 7.3 million were due to coronary disease then 6.2 million were due to stroke [4].
- A large number of people who die from heart diseases and mainly from heart disease and stroke, will increase to reach up to 23.3 to 30 million by 2030[5], CVDs are projected to remain the single leading cause of death and most heart diseases can be prevented by addressing risk factors.
- An 9.4 million deaths each year or 16.5% of all deaths can be attributed to high blood pressure[6]. It includes 51% of deaths due to stroke and 45% of deaths to coronary heart disease and other 4% of deaths due to other diseases and other problems of human body.
- It includes 16.5% of death due to this heart disease in overall India. in the every year death percentage are increasing the all over countries.
- Heart diseases affected for more men and wemon and responsible for more then 40% of all deaths due to united state.
- It includes 2.7% of death due to heartdisease in overall countries.

Hazards of heart disease:

Hazards for heart disease as follows

- Smoking.
- High blood pressure.
- Cholesterol.
- Obesity.
- Diabetes
- Brith control pills.

Existing systems are implemented with the SVM, Rough Set Techniques, Association Rule Mining used on the previous section to compare with this algorithm. Data mining could help in the identification or the prediction of high and low level of accuracy and provides the accurate results.Two techniques are implemented with classification tree, support vector mechanism to use on the previous paper. We proposed algorithms are implements with the classification tree, CN2 Rule, SOM and K-Means clustering to compared with this algorithm.

2. Related works

A heart disease is to built with the aid of data mining techniques like Support Vector Mechanism, Decision Tree was proposed to IJITEE (2012), they used on datasets in the heart disease. The heart disease using with datasets as for more than 250 data and above will be using the databases. There be an 8 to 14 attributes are used and such as age, sex, chest pain, cholesterol, fast blood pressure and etc…In the previous work on this paper, the result shows that SVM gave the lowest classification accuracy of 77.78% of higher and using the dataset taken from the UCI machine learning repository and the experimental result showed a correct classification accuracy of 77.56% with KNN. The analysis of a paper is represent with the K-Means Clustering, CN2 Rule, SOM, visualization, and unsupervised algorithms are used.

Python tool is used to classify the data and the data is evaluated using 2-fold cross validation and the results are compared, Python tool is a data mining suite built around GUI algorithms and the main purpose of this tool is given

to researchers and students an easy to use data mining software as to allowing to analysis either real or synthetic data.

To perform with the training dataset consists of 303 instances with 14 different attributes and data set is divided into two parts that is 80% of data is a training data and 20% of data is testing and the result, it is clear that classification accuracy of CN2 Rule, SOM algorithm is better compared to the algorithms. Heart disease and predicting system using data mining techniques as distribution diagram, CN2 Rule and SOM is implemented in [7], as a data source of data in a total number of records with the number of attributes from the heart disease in a database.CN2 Rule appears to be most effective as it has highest percentage of correct predictions with heart disease and classification trees, to be compared with the other models. In the K-Means clustering a large number of variables K-Means may be computationally faster than hierarchical clustering, and may be produce tighter cluster than hierarchical clustering and especially the cluster are globular [9].

3. Methodology

3.1. Classification Tree and Classification Algorithm

Heart disease or heart disease or coronary heart disease or ischemic heart disease [8] is a broad term that can refer to any condition that affected the heart disease. Today, they have several studies to applying different techniques to given problem and achieved high classification accuracies of 93.73% or higher and using the dataset taken from the UCI machine learning repository. Then, they have highest percentage of correct classification accuracy is also archived by 92.94%,86.14%,93.73% for the various algorithms produces these accuracy of high. The heart disease affects people of all income levels as[10,11,12].

3.2. The K-Means Algorithms:

It is a simple iterative method to partition a given dataset into a specified large sum of clusters; K this algorithm has been discovered by several researchers across the different disciplines. K-Means algorithms is one of the most commonly used partitioning clustering algorithms, as the "K" in it is name refer to the fact that the algorithm looks for a fixed number of clusters in terms of proximity of data points to each other. In this technique is iteration using with two-dimensional flow charts. The algorithm is usually handling mony more to type of elements. The points of corresponding to two elements of (x1, x2), the points correspond to n elements of the vectors (x1, x2… xn).

K-means algorithms

Where x, y is a set of two elements in the cluster and d(x,y) denote the min-distance between the two elements of x and x.

- ✓ Complete linkage clustering
  Its calculate the maximum distance between the large set of clusters. Formula as

$$d(x,y) = \max_{x\in X, y\in Y} d(x,y) \ldots\ldots (1)$$

where x, y is a set of two elements in the cluster and d(x, y) denote the max-distance between the two elements of x and x.

Its calculate the minimum distance between the large set of clusters. Formula as

$$d(x,y) = \min_{x\in X, y\in Y} d(x,y) \ldots\ldots (2)$$

where x, y is a set of two elements in the cluster and d(x, y) denote the min-distance between the two elements of x and x.

Euclidean distance

Euclidean distance are calculate the equation (3) will be used as

$$dis(x,y) = \sum_{i=1}^{d} \|x_i - c_i\| \ldots\ldots\ldots\ldots (3)$$

where x, y is a set of two elements in the cluster of dis(x,y) and i=1 is a counter value of cluster

Presently, the feature vectors are grouped into k means clusters using a selected distance measure such as Euclidean distance may be calculated .

The K-means and related algorithms gloss over the selection of K, there is no a priori reason to select a particular value and there is really an outermost loop to these algorithms that occurs during analysis rather than in the computer program [6]. The K-means clustering algorithm is the simplest and most commonly used algorithm it is

very sensitive to noise and original data points. Because a small number of data, such data can substantially influence the mean value [13].

3.3. SOM Algorithm

SOM or SOFM is a type of artificial neural network that is trained using unsupervised learning to produce a typical two dimensional(low-dimensional),discredited representation of the input space. SOMs useful for visualizing low dimensional views to high dimensional view of data to akin the multidimensional scaling data. In this algorithm is used for the analysis of heart disease using with data mining techniques in the unsupervised learning, it consists of components called nodes or neurons. An associated with each node is a weight vector of the same dimension as the input data vector and position in the vector space [14]. The nodes are arrangement is a two dimensional regular space in a hexagonal or rectangular grid, SOMs describes a mapping from a high dimensional input space to a lower dimensional map space. This type of network structure as related to feed forward networks where the nodes are visualized as being attached and this type of architecture is fundamentally different in arrangement and motivation.

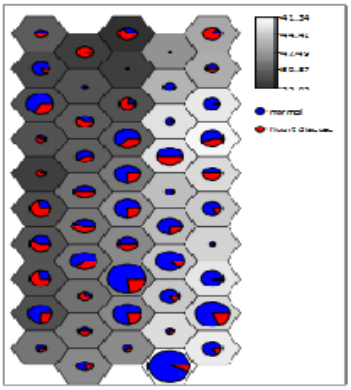

Figure 1: Distribution of SOMs visualizing map

3.4. CN2 Rule

CN2 rule to uses the likelihood ratio statistic (to developed by Clark and Niblett in year 1989) that measures the difference between the class probability distribution in the training data sets, the CN2 rule is used for two levels of controls are:

- Order list
- Unordered list

A default rule is to providing for majority class assignment as the final rule in the induced rule set. An ordered list of rules the procedure looks for the most accurate rule in the current set of training data and rule predicts the most frequent class in the set of covered data, CN2 rule finding the same rule again the all data sets are removed before a new iteration is started at a top level. Until all the data's are covered or no significant rule can be found, in unordered list of data control procedure is iterate and inducing rules for each class in turn of only covered data belonging to that class are removed, instead of removing all covered data[15].

CN2 rule are induced by the form 'if <complex> then predicting <class>' where <complex> as the definition of the CN2 rule, it is iterative fashion of each iteration searching for a complex that covers a large number of data's of single class C and few of other classes. It must be both predictive and reliable to determined by CN2's evaluation function and CN2 uses the simple method of replacing unknown values with the most commonly occurring value for that attribute in the training data.

Classification Tree Viewer:

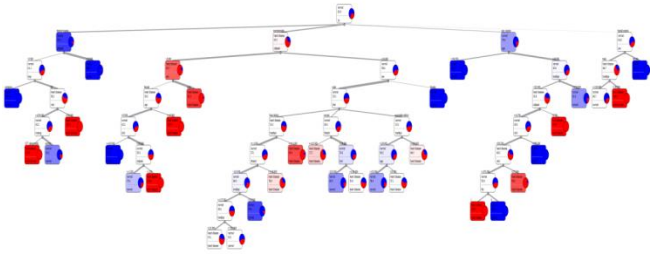

Figure2: to identification of heart disease using with classification tree.

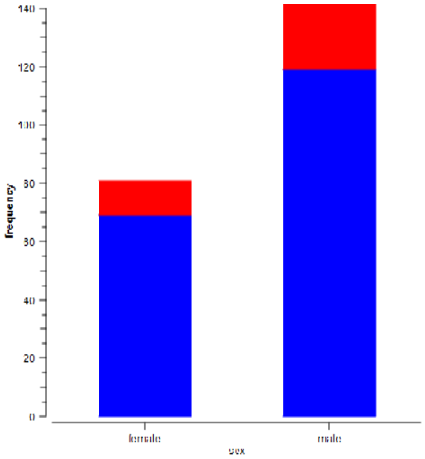

Figure3: attribute selection sex in the heart disease in a human.

3.5. Experiments and Classification Accuracy Table:

Algorithms	*Original data*	*Classification accuracy*
CN2	303	93.73
Classification tree	303	89.64
SOM	303	92.94

Table1: compared with various algorithms and distribute the classification accuracy table on heart disease datasets.

Table 1, we can calculate the classification accuracy in the original dataset as used with the heart disease in a data base and find out the accuracy equation as

$$AC = (\sum_{i=1}^{n} x_i) - ^{(\log_n n)}.(x_1 + x_2) \ldots\ldots\ldots\ldots\ldots (4).$$

wherex1, x2 is set elements in the positive prediction and negative prediction as the dataset in a heart disease and i=1 is a counter value of data in table and $x_i$is a original data in the table.

3.6. Classification accuracy chart

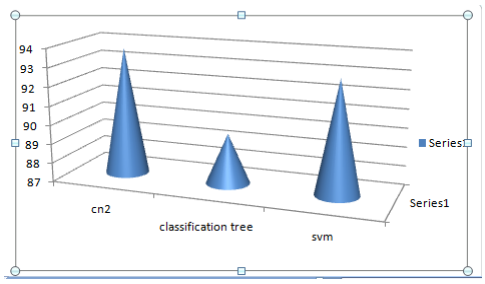

Figure 4: chart represented with classification accuracy of heart disease and various algorithms are implementd .

4. Conclusion:

In this paper , the goal was to design a predictive model for heart disease detection using data mining techniques from analysis the report for the classification accuracy among these data mining techniques has discussed, the result shows the difference in error rates. It is used with differences in different techniques as using to the classification tree and CN2 Rule to perform classification and more accurately than the other methods. The result of classification accuracy for 93.73% and missing error data as 6.27% or the incorrect data to the heart disease.

In the future work, we will try to increase the more accuracy for heart disease patient by increasing the different parameters and will be compared with different algorithms using with data mining techniques.

References

[1] data mining concept:-cgi.di.uao.gr/~rouvas/research/white.html.

[2] Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[3] Global status report on noncommunicabledisaeses 2010. Geneva, World Health Organization, 2011.

[4] Global atlas on heart disease prevention and control. Geneva, World Health Organization, 2011.

[5] Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med, 2006, 3(11):e442.

[6] Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet, 2012, 380(9859):2224–2260.

[7] sellappapalaniapparafiahswang,intelligent heart disease prediction system using data mining techniques ,IJCSNS,vol 8 no 8 aug 2011.

[8] BalaSundar V, Bharathiar, ―Development of a Data Clustering Algorithm for Predicting Heart‖ International Journal of Computer Applications (0975 – 888) Volume 48– No.7, June 2012

[9] http://heart-disease.emedtv.com/ coronary-artery-disease/coronary-artery-disease.html

[10] R. Gupta, V. P. Gupta, and N. S. Ahluwalia, ―Educational status, coronary heart disease, and coronary risk factor prevalence in a rural population of India‖, BMJ. pp 1332–1336, 19 November 2012.

[11] A Mathavan, MD, A Chockalingam, PhD, S Chockalingam, BSc, B Bilchik, MD, and V Saini, MD, ―Madurai Area Physicians Heart Health Evaluation Survey (MAPCHES) – an alarming status‖, The Canadian Journal of Cardiology; 25(5): 303–308, May 2009.

[12] Rajeswari K, Vaithiyanathan V, P. Amirtharaj ―Application of Decision Tree Classifiers in Diagnosing Heart Disease using Demographic Data‖ American Journal of Scientific research ISSN 2301-2005 pp. 77-82 EuroJournals Publishing, 2012.

[13] Mrs. Bharati M. Ramageri, ―Data Mining Techniques And Applications‖,Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305

[14] ^ **a b** Ultsch, Alfred (2003); U*-Matrix: A tool to visualize clusters in high dimensional data, Department of Computer Science, University of Marburg, Technical Report Nr. 36:1-12

[15] D. Gamberger, N. Lavraˇc, F. ˇ Zelezn´y, and J. Tolar. Induction of comprehensible models for geneexpression datasets by subgroup discovery methodology. JrBiomedical Informatics, 37(5):269–284, 2004