Volume 1, No. 3, May 2012 ISSN – 2278-1080
The International Journal of Computer Science & Applications (TIJCSA)
RESEARCH PAPER
Available Online at http://www.journalofcomputerscience.com/

Predicting the Heart Attack Symptoms using Biomedical Data Mining Techniques

V.V.Jaya Rama krishniah
Associate Professor,
ASN. College, Tenali
jkvemula@yahoo.com

D.V.Chandra Sekar
Associate Professor,
TJPS College, Guntur
chand.info@gmail.com

Dr.K.Ramchand H Rao
Professor,
ASN Womens Engineering College, Nalapadu, Tenali
ramkolasani@gmail.com

Abstract

The diagnosis of heart disease is a significant and tedious task in medicine. The healthcare industry gathers enormous amounts of heart disease data that regrettably, are not “mined” to determine concealed information for effective decision making by healthcare practitioners. The term Heart disease encompasses the diverse diseases that affect the heart. Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The reduction of blood and oxygen supply to the heart leads to heart disease. In this paper the data classification is based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. Tanagra tool is used to classify the data and the data is evaluated using entropy based cross validations and partitioned techniques and the results are compared.

Keywords: K- mean, Euclidian distance, Nearest Neighbor, Entropy based Mean Clustering.

1. INTRODUCTION

The term heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels. It is intended to deal only with the condition commonly called "Heart Attack" and the factors, which lead to such condition. Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term -

cardiovascular disease "includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death. Narrowing of the coronary arteries results in the reduction of blood and oxygen supply to the heart and leads to the Coronary heart disease (CHD). Myocardial infarctions, generally known as a heart attacks, and angina pectoris, or chest pain are encompassed in the CHD. A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack. Chest pains arise when the blood received by the heart muscles is inadequate. High blood pressure, coronary artery disease, alular heart disease, stroke, or rheumatic fever/rheumatic heart disease are the various forms of cardiovascular disease

1) *Early Signs Of Heart Disease*

1. Dizzy spell or fainting fits.
2. Discomfort following meals, especially if long continued.
3. Shortness of breath, after slight exertion.
4. Fatigue without otherwise explained origin.
5. Pain or tightness in the chest a common sign of coronary insufficiency is usually constrictive in nature and is located behind the chest bone with
6. Radiation into the arms or a sense of numbness or a severe pain in the centre of the chest.
7. Palpitation

2. EXTRACTION OF HEART DISEASE DATAWAREHOUSE

The heart disease data warehouse contains the screening the data of heart patients. Initially, the data warehouse is pre-processed to make the mining process more efficient. In this paper Tanagra tool is used to compare the performance accuracy of data mining algorithms for diagnosis of heart disease dataset. The pre-processed data warehouse is then classified using Tanagra tool. The feature selection in the tool describes the attribute status of the data present in the heart disease. Using unsupervised machine learning algorithm such as k-nn, K-Mean and Entropy based clustering and the result are compared. Tanagra is a collection of machine learning algorithms for data mining tasks. The algorithms can be applied directly to a dataset. Tanagra contains tools for data classification, statistics, clustering, supervised learning, meta-supervised learning and visualization. It is also well suited for developing new machine learning schemes. This paper concentrates on functional algorithms like K-NN, K-Mean and Entropy based Clustering.

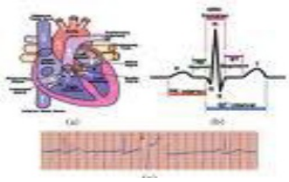


Figure 1: Diagnosis of heart disease

3. TANAGRA

Tanagra is a data mining suite build around graphical user interface. Tanagra is particularly strong in statistics, offering a wide range of uni- and multivariate parametric and nonparametric tests. Equally impressive is its list of feature selection techniques. Together with a compilation of standard machine learning techniques, it also includes correspondence analysis, principal component analysis, and the partial least squares methods. Tanagra is more powerful, it contains some supervised learning but also other paradigms such as clustering, supervised learning, meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain , and allowing to analyze either real or synthetic data. Tanagra can be considered as a pedagogical tool for learning programming techniques. Tanagra is a wide set of data sources, direct access to data warehouses and databases, data cleansing, interactive utilization.

1) *Clustering*

The Clustering is based on unsupervised algorithms. Clustering is done to know the exactly how data is being classified. Tanagra includes support for arcing, boosting, and bagging classifiers to make clustering. These algorithms in general operate on a clustering algorithms like K-Mean and run it multiple times manipulating algorithm parameters or input data weight to increase the accuracy of the cluster seed. Two learning performance evaluators are included with Tanagra. The first simply splits a dataset into training and test data, while the second performs cross-validations between the centers of the defined number of clusters. Evaluation is usually described by accuracy, error, precision and recall rates.

2) *Manifold machine learning algorithm*

The main motivation for different unsupervised machine learning algorithms is accuracy improvement. Different algorithms use different rule for generalizing different representations of the knowledge. Therefore, they tend to error on different parts of the instance space. The combined use of different algorithms could lead to the correction of the individual uncorrelated errors. As a result the error rate and time taken to develop the algorithm is compared with different algorithm.

3) *Algorithm selection*

Algorithm is selected by evaluating each supervised machine learning algorithms by using supervised learning assessment on the training set and selects the best one for application on the test set. Although this method is simple, it has been found to be highly effective and comparable to other methods. Several methods are proposed for machine learning domain. The overall cross validation performance of each algorithm is evaluated.

The selection of algorithms is based on their performance, but not around the test dataset itself, and also comprising the predictions of the classification models on the test instance. Training data are produced by recording the predictions of each algorithm, using the full training data both for training and for testing. Performance is determined by running 10-fold cross-validations and averaging the evaluations for each training dataset. Several approaches have been proposed for the characterization of learning domain. the performance of each algorithm on the data attribute is recorded. The algorithms are ranked according to their performance of the error rate.

This paper deals with K-mean, Nearest Neighbor, and Entropy based Mean clustering algorithm. Experimental setup is discussed the training data set consists 14 different attributes. The information was extracted from the UCI Machine learning web site. The performance analysis is done among these algorithms based on the accuracy and time taken to build the model. Tanagra Software is used for K Mean and Nearest Neighbor Procedures and Entropy based mean clustering is developed using Advanced Java.

4. ALGORITHM USED

The *k*-nearest neighbor's algorithm (*k*-NN) is a method for classifying objects based on closest training data in the feature space. k-NN is a type of instance-based learning. The function is only approximated locally and all computation is deferred until classification. The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its *k* nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data .Nearest neighbor rules in effect compute the decision boundary in an implicit manner. It is also possible to compute the decision boundary itself explicitly, and to do so in an efficient manner so that the computational complexity is a function of the boundary complexity. The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good *k* can be selected by various heuristic techniques, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when $k = 1$) is called the nearest neighbor algorithm. The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. In binary (two class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes. Using an appropriate nearest neighbor search algorithm makes *k*-NN computationally tractable even for large data sets. The nearest neighbor algorithm has some strong consistency results. As the amount of data approaches infinity, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data. *k*-nearest neighbor is guaranteed to approach the Bayes error rate, for some value of *k* (where *k* increases as a function of the number of data points). The k-means is the simplest and most commonly used clustering algorithm. The simplicity is due to the use of squared error as the stopping criteria, which tends to work well with isolated and compact clusters. Its time

complexity depends on the number of data points to be clustered and the number of iteration. The K mean algorithm works on the Euclidian Distance Method, is initialized from some random or approximate solution. Each iteration assigns each point to its nearest cluster and then points belonging to the same cluster are averaged to get new cluster centroids. Each iteration successively improves cluster centroids until they become stable. The Entropy based Mean Clustering algorithm (EMC) is extension to the K mean algorithm, reduces the number of iterations during the clustering process. It works on three phases. In the first phase it computes the min points of the each seed (element or item) in the data set and then arranges the seed elements in the order of their seed entropy (For example 1-10,2-5,3-9,4-6,5-1, then it arranges the data as 1,3,4,2,5 .i.e. data arranged ascending order of the entropy). In the second phase, it makes the candidate set, this candidate set is unique in nature, i.e it does not consisting of duplicated elements. In the third phase the clustering was applied on the Euclidian distances, and remaining elements, which were not in candidate sets were placed in according to the native elements, were resided.

5. EXPERIMENTAL SETUP

The data mining method used to build the model based on clustering. The data analysis is processed using Tanagra data mining tool for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analysed.

1) *Description of dataset*

The dataset contains the following attributes:

1) id: patient identification number

2) age: age in year,

3) sex: sex (1 = male; 0 = female),

4) painloc: chest pain location (1 = substernal; 0 = otherwise),

5) painexer (1 = provoked by exertion; 0 = otherwise),

6) relrest (1 = relieved after rest; 0 = otherwise),

7) cp: chest pain type

- Value 1: typical angina
- Value 2: atypical angina
- Value 3: non-anginal pain
- Value 4: asymptomatic

8) trestbps: resting blood pressure

9) chol: serum cholestoral

10) famhist: family history of coronary artery disease (1 = yes; 0 = no)

11) restecg: resting electrocardiographic results

- Value 0: normal
- Value 1: having ST-T wave abnormality (T wave inversions and/or ST
- elevation or depression of > 0.05 mV)
- Value 2: showing probable or definite left ventricular hypertrophy
- by Estes' criteria

12) ekgmo (month of exercise ECG reading)

13) thaldur: duration of exercise test in minutes

14) thalach: maximum heart rate achieved

15) thalrest: resting heart rate

16) num: diagnosis of heart disease (angiographic disease status)

- Value 0: < 50% diameter narrowing
- Value 1: > 50% diameter narrowing

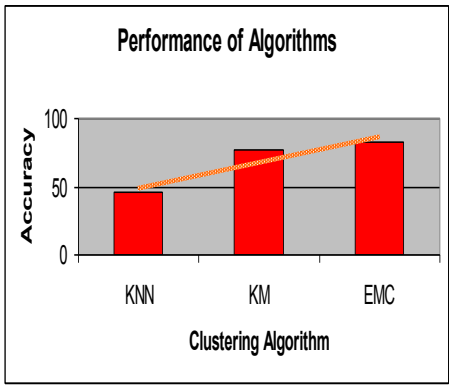
17) (in any major vessel: attributes 59 through 68 are vessels)

2) Performance study of Algorithms

Performance can be determined based on the evaluation time of calculation. Comparison is made among the clustering algorithms. The following Table 1 consists of secondary values of different clustering algorithms. According to these values the accuracy is calculated and analyzed. Each one has a distinct value.. Because it takes only some time to calculate the accuracy than other algorithms. To evaluate performance of the clustering, we used the demission (attribute) called restecg's St-t- abnormal property

Algorithm	Accuracy	Time
KNN	45.67	1200ms
KM	76.90	783ms
EMC	82.90	699ms

Table 1. Performance Study Of Algorithm



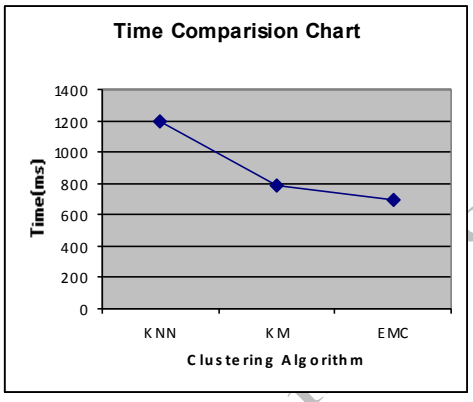


Fig1: Performance of the algorithm

Fig 2: Time duration for each algorithm

From the Fig1 and Fig 2, we notice that the performance and time taken to make the clustering is lesser for Entropy based mean clustering mechanism.

6. CONCLUSION

Data mining in health care management is unlike the other fields owing to the fact that the data present are heterogeneous and that certain ethical, legal, and social constraints apply to private medical information. Health care related data are voluminous in nature and they arrive from diverse sources all of them not entirely appropriate in structure or quality. These days, the exploitation of knowledge and experience of numerous specialists and clinical screening data of patients gathered in a database during the diagnosis procedure, has been widely recognized. This paper deals with the results in the field of clustering based on k-Nearest Neighbor, K Mean and Entropy based mean clustering algorithms, and on the whole performance made known Entropy based mean is the best compact time for processing dataset and shows better performance in accuracy prediction. The time taken to run the data for result is fast when compared to other algorithms. It shows the enhanced performance according to its attribute. Attributes are fully classified by this algorithm and it gives 82.90% of accurate result.

REFERENCE

1) Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108 ,1999.

2) Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo 1993.

3) "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.

4) M. Chen, J. Han, and P. S. Yu. Data Mining: An Overview from Database Perspective. IEEE Trans. Knowl. Dat. Eng., vol: 8, no:6, pp: 866-883, 1996.

5) R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases", In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, August 29-September 1994.

6) Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review , Vol. 8, No. 1 (January - June 2007).

7) Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation atUniversity of Tokyo, 2004.

8) Franck Le Duff , Cristian Munteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics ,107(Pt 2):1256-9, 2004.

9) Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, Karsten Sternickel, Lijuan Zhu, "Using Efficient Supanova Kernel For Heart Disease Diagnosis", proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16, pp:305-310, 2006.

10) Agrawal, R., Imielinski, T. and Swami, A, _Mining association rules between sets of items in large databases', Proc. ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'93),Washington, DC, July, pp.45–51, 1993.

11) "Heart Disease" from http://chineseschool. netfirms.com/heart-disease-causes.html

12) "Heart disease" from http://en.wikipedia.org/wiki/Heart_disease.

13) Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee,and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", springer,Vol:345 , pp 721-727, 2006.

14) Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

15) Andreeva P., M. Dimitrova and A. Gegov, Information Representation in Cardiological Knowledge Based System,SAER'06, pp: 23-25 Sept, 2006.