TO ASCERTAIN AND CONTEMPLATE CARDIO-VASCULAR PROGNOSTICATION IN DATA MINING

Vinoth Haribabu[1], Surya Veeraraghavalu[2]

[1]Research Scholar, Department of computer science, Government Arts College for Men, Nandanam, Chennai -600 035

[2]Assistant Professor, Department of computer science, Government Arts College for Men, Nandanam, Chennai -600 035

Abstract *-- Cardiovascular disease is one of the most fatal conditions in the present world. Statistical information shows the harmful of Cardiovascular ailment by uncovering the level of deaths overall caused because of heart assaults. Hence there is an implicit necessity to predict the condition at the earliest. However, with the increasing complexity of the data especially in biomedical and healthcare communities, accurate analysis of medical data needs to be done to get accurate results for disease detection. By using the patient's medicinal records, another system is proposed to anticipate the chances of an individual contracting heart assault. Properties, for example, age, pulse, thickness of the corridor, and so forth are utilized to foresee danger of heart assault in an individual. In the proposed work, the prediction analysis is applied to predict data according to the biomedical dataset. In this work, the k-means clustering algorithm and SVM (support vector machine) classifier-based prediction analysis technique will be used for clustering and classification of the input data. To increase the accuracy of prediction analysis, the back-propagation algorithm is proposed to be applied with the k-means clustering algorithm to cluster the data.*

Keywords-- *ANFIS - Adaptive Neuro-Fuzzy Inference System, CANFIS – Co-Active ANFIS, HDD - Heart Disease Data, KNN – K Nearest Neighbor, SVM – Support Vector Machine*

1. INTRODUCTION

The prediction analysis is the technique which can predict the future possibilities from the existing data. The prediction analysis techniques are based on clustering and classification. In base paper medical data is analyzed to predict the regional diseases. The process of prediction analysis is divided into two phases which are clustering and classification. In the existing work, k-mean clustering is applied to cluster head and output of clustering is given as input to SVM classifier for the classification. In k-mean algorithm, the centered points are calculated by taking arithmetic mean of the whole data set. The points which have similar value is clustered in one cluster and other in the second cluster. In the k-mean clustering algorithm some points remained un-clustered which reduces accuracy. The Dataset of heart disease is collected from the UCI repository which contains 14 attributes. To apply k-mean clustering on the dataset whole dataset is taken as input with all the instances. The k-mean clustering divided whole dataset into similar type of groups. The output of k-mean clustering will be given as input to SVM classifier which can classify data based on hyper planes. In this research work, the KNN classifier will be applied on the place of SVM classifier for the prediction analysis. The output of clustering will be given as input for the classification which increases accuracy of prediction analysis. Through this paper, we reach an inference that the accuracy of the proposed model is better than existing strategies.

2. DATA MINING

Data mining is the way toward discovering patterns in substantial data sets including strategies at the intersection of machine learning, statistics, and database systems.[1] Data mining is an interdisciplinary subfield of computer science and statistics with a general objective to extract information (with astute techniques) from an data set and change the data into an intelligible structure for further use.[2] Data mining is the analysis step of the "learning disclosure in databases" process, or KDD.[3] Aside from the raw analysis step, it likewise includes database and data management aspects, information pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, representation, and online updating.[1] The distinction between information investigation and information mining is that information investigation is to abridge the history, for example, breaking down the adequacy of an advertising effort, interestingly, data mining centers around utilizing explicit machine learning and factual models to anticipate the future and find the examples among data.

2.1 Techniques used

2.1.1 Classification

Classification is an exemplary data mining technique based on machine learning. Classification is used to classify each item in a set of data into one of established set of classes or groups. Classification technique makes use of mathematical techniques, for example, decision trees, linear programming, neural system and statistics.

2.1.2 Clustering

Clustering is the gathering of a specific set of objects based on their characteristics, aggregating them as indicated by their similarities. As per Data mining, this technique segments the information actualizing a specific join algorithm, most appropriate for the ideal information analysis.

This clustering analysis permits an object not to be a part of a cluster, or strictly belong to it, calling this kind of grouping hard apportioning. Then again, soft partitioning states that each object has a place with a cluster in a decided degree. Progressively explicit divisions can be conceivable to make like objects belonging to multiple clusters, to drive an object to participate in just a single cluster or even develop hierarchical trees on group relationships.

2.1.3 Association

Association is the best-known data mining technique. In association, a pattern is exposed based on a relationship of a specific item on other items in a similar task. For instance, the association technique is utilized in coronary illness prediction as it says to us the relationship of dissimilar attributes utilized for analysis and deal with the patient with all the hazard factor which are essential for prediction of disease.

2.1.4 Prediction

The prediction as it names backhanded is one of a data mining technique that finds relationship between independent variables and relationship among reliant and autonomous factors. For instance, prediction analysis technique can be utilized in deal to foresee benefit for the future if consider sale is an independent variable, profit could be a dependent variable. At that point dependent on the historical sale and profit data and can draw a fixed regression curve that is utilized for revenue driven forecast.

Table 1. Data Mining Tasks and Intelligent Techniques

S.No	Data Mining task	Data Mining Algorithm and Technique
1	Classification	Decision Trees, Rule-based, Neural Networks, Naive Bayes and Bayesian Belief Networks, Support Vector Machines, Genetic Algorithms
2	Clustering	K-Means
3	Regression and Prediction	Support Vector Machines, Decision Trees, Rule induction, NN
4	Association and Link Analysis (finding correlation between items in a dataset)	Association Rules Mining (ARM)

3. ALGORITHM AND TOOL USED

3.1. SUPPORT VECTOR MACHINE

SVM is a well-known classifier that is utilized in regression, classification and general pattern recognition [4]. The initial form of SVM is a binary classifier where the yield of the learned function is either positive or negative. There is no need to include any earlier knowledge. At the point when there is the high dimension of input space at that point utilizing Kernel methods it offers better outcomes. SVMs does the mapping from input space to feature space to support nonlinear classification problems. The kernel trick is useful for doing this by permitting the absence of the exact formulation of mapping function which could cause the issue of the curse of dimensionality. Geometric representation is the best classification function within this methodology. An isolating hyperplane that goes through the middle of two classes is journalist to the linear classification function in the event of a linearly separable dataset.

3.2. K-NEAREST NEIGHBOR

The learning performed utilizing analogy is the base of KNN classifiers. With the assistance of n-dimensional numeric attributes, the description of the training samples is done. A point inside the n-dimensional space is represented by each sample. Along these lines, inside the n-dimensional pattern space, all the training samples are recorded. The pattern space for k training samples which are closest to obscure examples is looked by KNN classifier for the situation when an obscure example is given. The Euclidean samples helps in characterizing the closeness of samples. As all the attributes is deferred to that time span, the speed of characterization turns out to be less. Each attribute is assigned with equal weight by nearest neighbor classifiers which are impractical in decision tree induction and backpropagation. On the off chance that there are a few irrelevant attributes inside the data, confusion may be generated here. To give a prediction, the nearest neighbor classifiers can likewise be used with such that for a given unknown sample, the real-valued prediction can be returned [5]. The normal estimations of genuine qualities that are related with k-closest neighbors are returned here by this classifier. Among all other machine learning algorithms, the KNN is the least complex one. In view of the larger part votes of the neighbors, an article can be classified.

3.3 ANACONDA

Anaconda is a freemium open source distribution of Python and R programming for large scale information processing, predictive analytics, and logical computing, that intends to disentangle package management and deployment. Its package management framework is conda.

4. LITERATURE SURVEY

We took few previous papers for literature survey to arrive towards a conclusion on how to further enhance the system in terms of accuracy.

4.1 Latha Parthiban and R.Subramanian - 2007

In paper titled Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm author updated that “Heart disease (HD) is a major cause of morbidity and mortality in the modern society. Medical diagnosis is an important but complicated task that should be performed accurately and efficiently, and its automation would be very useful. All doctors are unfortunately not equally skilled in every sub specialty and they are in many places a scarce resource. A system for automated medical diagnosis would enhance medical care and reduce costs. In this paper, a new approach based on coactive neuro-fuzzy inference system (CANFIS) was presented for prediction of heart disease. The proposed CANFIS model combined the neural network adaptive capabilities and the fuzzy logic qualitative approach which is then integrated with genetic algorithm to diagnose the presence of the disease. The performances of the CANFIS model were evaluated in terms of training performances and classification accuracies and the results showed that the proposed CANFIS model has great potential in predicting the heart disease.”

4.2 Jabbar Akhil and Bulusu Deekshatulu – 2012

In paper titled Heart Disease Prediction System using Associative Classification and Genetic Algorithm author updated “Associative classification is a recent and rewarding technique which integrates association rule mining and classification to a model for prediction and achieves maximum accuracy. Associative classifiers are especially fit to applications where maximum accuracy is desired to a model for prediction. There are many domains such as medical where the maximum accuracy of the model is desired.

Heart disease is a single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. Mortality data from the registrar general of India shows that heart disease are a major cause of death in India, and in Andhra Pradesh coronary heart disease cause about 30%of deaths in rural areas. Hence there is a need to develop a decision support system for predicting heart disease of a patient. In this paper we propose efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. Experimental Results show that most of the classifier rules help in the best prediction of heart disease which even helps doctors in their diagnosis decisions."

4.3 V. Krishnaiah, G. Narasimha – 2016

In paper titled Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review authors updated "The Healthcare trade usually clinical diagnosis is ended typically by doctor's knowledge and practice. Computer Aided Decision Support System plays a major task in medical field. Data mining provides the methodology and technology to alter these mounds of data into useful information for decision making. By using data mining techniques, it takes less time for the prediction of the disease with more accuracy. Among the increasing research on heart disease predicting system, it has happened to significant to categories the research outcomes and gives readers with an outline of the existing heart disease prediction techniques in each category. Data mining tools can answer trade questions that conventionally in use much time overriding to decide. In this paper we study different papers in which one or more algorithms of data mining used for the prediction of heart disease. As of the study it is observed that Fuzzy Intelligent Techniques increase the accuracy of the heart disease prediction system. The generally used techniques for Heart Disease Prediction and their complexities are summarized in this paper.

4.4 Gaurav Meena, Pradeep Singh Chauhan – 2017

In paper titled Empirical Study on Classification of Heart Disease Dataset its Prediction and Mining authors updated "Despite the significance of data mining techniques to Heart Disease Data (HDD), there be a short of comprehensive literature review and a classification for it. This is the academic literature review of the application of data mining techniques to HDD (Heart Disease Data). It provides an academic database of literature amid the periods of 20062016 and proposes a classification scheme to classify the Mining techniques. The analysis and classification process was autonomously verified. Findings of this paper indicate that the research area of Heart Disease received most research attention. Classification and Association models are the two most commonly used models for data mining in Heart Disease Data. Our study provides a path to provide future research ideas and make easy creation about the application of data mining techniques in HDD."

5. DISADVANTAGES OF EXISTING SYSTEM:

We took 2017 paper as base paper and found few disadvantages. Below are the disadvantages of existing system:

- It is computationally expensive to find the k nearest neighbors when the dataset is very large
- Central point is not calculated accurately due to which some point remained un-clustered
- Accuracy of prediction was reduced

6. PROPOSED SYSTEM:

The k-mean clustering is the clustering technique in which similar and dissimilar data is clustered together based on their similarity. In the k-mean clustering, the dataset is considered and from that dataset arithmetic mean is calculated which will be the central point of the dataset. The Euclidian distance from the central point is calculated and points which are similar and dissimilar are clustered into different clusters. The KNN classifier will be applied which can classify the data into certain classes. The certain classes mean the entries which has the heart disease, and which do not have heart disease.

Below steps are proposed to implement in our work:

- Input data values
- Division of Input data
- Classify data

The proposed technique will be implemented in python along with ANACONDA tool and results will be analyzed in terms of accuracy, execution time.

6.1 ADVANTAGES OF PROPOSED SYSTEM

Following are the various benefits of this proposed work:

- The proposed system gives high accuracy for the heart disease prediction which leads to improve system efficient of disease prediction
- The proposed algorithm also improves various other parameters like precision, f- measure and execution time for heart disease prediction

6.2 RESEARCH METHODOLOGY

The prediction analysis technique is used to predict the situations according to the input dataset. The prediction analysis requires two phases. In the first phase, the k-mean clustering is applied which will cluster the similar and dissimilar type of data. In the principal stage, the k-mean clustering is connected which will cluster the comparative and unique kind of data. In the second stage, the SVM classifier is applied which will classify the information. The k-mean clustering comprises of three stages. The initial step, the arithmetic mean of the entire dataset is determined which is taken as the central point. The second step, Euclidean distance is determined for all points from the central points. Finally, the data will be clustered according to their similarity. The clustered data will be given as input to the SVM classifier for the classification. The data classification quality depends upon the cluster quality. In this thesis, the k-mean clustering algorithm will improve the quality of cluster which increases the quality of classification. Backpropagation algorithm is used to calculate the Euclidean.

6.3 STEPS USED TO IMPLEMENT

1. **Input Data Values:** -The output of k-mean clustering is given as input to the KNN classifier for the classification. The KNN classifier can classify data into certain set of classes like heart disease and non- heart disease

2. **Division of Input Data:** -In the second step, whole data will be divided into training and test set. The training set will be 60 percent of the whole data and test set will be the 40 percent of the whole data.

3. **Classify Data:** - The training and test set will be given as input for the classification. The function called kNN is applied which take input training and test set for the classification. The KNN classifier can draw hyper plane which can divided data into heart disease and non-heart disease classes.

7. RESULTS AND CONCLUSION

In this paper discussed an investigation of various data mining techniques that can be utilized in automated coronary illness prediction systems. The analysis demonstrates that distinctive technologies are utilized in all the papers with taking diverse number of attributes reached their outcomes distinctive accuracy relies upon tools utilized for implementation. Even though applying data mining systems to help health care professionals in the determination of coronary illness is having different victories. This paper gives a quick and thoughtful of various prediction models in data mining and finds most prominent model for further work. This work can be improved by increasing the number of attributes for the existing system of our past work. We can have an increasingly number of records to give better exactness to the framework in foreseeing and diagnosing the patients of coronary illness.

REFERENCES

[1] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27

[2] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09

[3] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Archived from the original on 2009-11-10. Retrieved 2012-08-07

[4] Chen, Z.X., 2009. Shixiong, "K-means Clustering Algorithm with improved Initial Center," in Second international Workshop on Knowledge Discovery and Data Mining, Moscow

[5]https://www.researchgate.net/publication/328512223_K-means_based_SVM_for_Prediction_Analysis

[6] A. Taneja, "Heart disease prediction system using data mining techniques," Oriental Journal of Computer science and technology, vol. 6, pp. 457-466, 2013.

[7] P. Cortez, "Data mining with neural networks and support vector machines using the R/rminer tool," in Industrial Conference on Data Mining, 2010, pp. 572-583.

[8] C. Velu and K. Kashwan, "Visual data mining techniques for classification of diabetic patients," in Advance Computing Conference (IACC), 2013 IEEE 3rd International, 2013, pp. 1070-1075.

[9] N. P. Waghulde and N. P. Patil, "Genetic neural approach for heart disease prediction," *International Journal of Advanced Computer Research*, vol. 4, p. 778, 2014.

[10] V. C. Osamor, E. F. Adebiyi, J. O. Oyelade, and S. Doumbia, "Reducing the Time Requirement of k-means Algorithm," *PLoS One*, vol. 7, p. e49946, 2012.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273-297, 1995.

[12] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, pp. 293-300, 1999.

[13] A. Rauf, S. M. Sheeba, S. Khusro, and H. Javed, "Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity," *Middle-East Journal of Scientific Research*, vol. 12, pp. 959-963,2012.