Research Paper Computer Science

Heart Disease Analysis And Prediction Using Various Classification Models – A Survey

P.DEEPIKA ASST.PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, SASURIE COLLEGE OF ARTS AND SCIENCE, VIJAYAMANGALAM.

P.VINOTHINI Student, Department of Information Technology,Kathir College of Engineering,Coimbatore.

ABSTRACT

Data mining is the process of discovering or extracting new patterns from large datasets involving methods from statistics and artificial intelligence. Today health care industry generates large amount of complex data about patients, hospitals, resources, diseases and their records etc. This large amount of data is the essential resources to be processed and analyzed for knowledge extraction that supports for cost saving and decision making. Heart disease prediction system can assist medical professionals in predicting heart disease statuses based on the clinical data of the patient.

KEYWORDS Data Mining, Knowledge Extraction, Prediction.

INTRODUCTION

Data mining refers to extracting or mining knowledge from large amount of data. There are several tools are available for data mining. Data mining in medical science is critical and is more sensitive than other domains because of its complexity of nature. On the other hand the significance of data mining in medical science can play a vital role if it is utilized for prediction and decision making. Today the health care industry generates large amount of complex data about patients, disease diagnosis, electronic patient records or medical devices. Data mining provides a set of tools and techniques that can be applied for knowledge extraction for decision making. Data mining in health care is an rising field of high importance for prediction and a deeper understanding of medical data. Classification is one of the important data mining techniques with broad applications to classify the various kinds of data used in every field of our life. Classification and prediction are the techniques used to make out important data classes and predict probable trend. The term heart disease is related to all diverse diseases affecting the heart.

CLASSIFICATION

Classification is commonly used data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Classification is a classical data mining technique based on machine learning. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. A two step process are involved in classification.

That are:

- Model construction
- Model usage

Model construction describes a set of predetermined classes. Each sample is assumed to belong to a predefined class as determined by the class label attribute. The set of samples used for model construction: training set. The model is represented as classification rules, decision trees or mathematical formula.

Model usage is used for classifying future and unknown objects. Estimate accuracy of the model. Accuracy rate is used to show the percentage of test set samples that are correctly classified by the prescribed model. Test set should be an independent of training set, otherwise over-fitting will occur.

The classifier training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

TYPES OF CLASSIFICATION MODELS

- Classification by decision tree
- Bayesian Classification
- Neural Networks
- Support vector machine
- Classification based on association

HEART DISEASE

Heart disease is one of the major cause of death all over the world in the past ten years. Cardiovascular disease is caused by disorder of the heart and blood vessels and includes coronary heart disease and heart failure. The major reason of heart disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol.

To prevent and identification of these heart diseases different techniques of data mining is used through this easily find out heart related diseases and this is the aim of this research work.

FOCUS ON THE SURVEY

Chaurasia and Pal conducted study on the prediction of heart attack risk levels from the heart disease database. The prediction of heart diseases significantly uses 11 important attributes, with basic data mining technique like Naïve Bayes, J48 decision tree and Bagging approaches. The outcome shows that bagging techniques performance is more accurate than Bayesian classification and J48. The results shows that the bagging prediction system is capable of predicting the heart attack effectively [1].

JAM is a powerful and portable agent-based distributed data mining [2] system that employs meta learning techniques to integrate a number of independent classifiers (models) derived in parallel from independent and (possibly) inherently distributed databases. Although meta-learning promotes scalability and accuracy in a simple and straightforward manner, brute force meta-learning techniques can result in large, redundant, inefficient and sometimes inaccurate meta-classifier hierarchies. In this paper they explore several methods for evaluating classifiers and composing meta-classifiers, they expose their limitations and they demonstrate that meta-learning combined with certain pruning methods has the potential to achieve similar or even better performance results in a much more cost effective manner.

Random forests are one of the most successful ensemble methods[3][10] which exhibits performance on the level of boosting and support vector machines. The method is fast, robust to noise, does not over fit and offers possibilities for explanation and visualization of its output. They investigate some possibilities to increase strength or decrease correlation of individual trees in the forest. Using several attribute evaluation measures instead of just one gives promising results. On the other hand replacement of ordinary voting with voting weighted with margin achieved on most similar instances gives improvements which are statistically highly significant over several data sets.

Heart disease or coronary diseases are a broad term that can refer to any condition that affects the heart. Karolis et al has developed a data mining system for the assessment of heart related risk factors targeting in reduction of CHD events using risk factors. The analysis was carried out using C4.5 decision tree algorithm for predicting heart disease. The highest percentage of classification was 66%, 75% and 75% for the myocardial infarction Percutaneous Coronary Intervention and Coronary Artery Bypass Graft Surgery[4].

G.Subbalahshmi et. al. Proposed a work to predict heart disease using Naive Bayes[5]. The main objective of this research work is to develop a decision support in heart disease prediction using Naive Bayes algorithm. This system extracts hidden useful information from the heart disease database. This model may possibly answer difficult queries, each one with own potency with respect to ease of model analysis, access to complete information and accurateness. This model can be further enhanced and expanded by incorporating other data mining techniques.

S.Vijiyarani et. al. A research work to predict the heart disease using classification tree technique[6]. The classification tree algorithms used and tested in this work are Decision Stump, Random Forest and LMT Tree algorithm. The objective of this work was to compare the outcomes of the performance of different classification techniques for a heart disease dataset. This work was done with WEKA tool.

Deepali et. al. Proposed a research work of diagnosing heart disease using data mining algorithm[7]. Datasets used for this research work was taken from UCI centre for machine learning and intelligent system. The proposed work showed that how information gain method, feature selection technique can be used in collaboration with adaptive neuro fuzzy inference systems in diagnosing new patient cases. This combination created a new approach for the heart disease by reducing the number of features to the optimal number using the information gain and then applied the new dataset to ANFIS. This study found that the accuracy for the proposed approach was 98.24% compared with other methods.

CONCLUSION

In this survey paper, different data mining techniques are used in the field of Heart Disease is discussed. Mainly five different classification algorithms namely Classification by decision tree, Bayesian Classification, Neural Networks, Support vector machine, Classification based on association are focused in the field of Heart Disease Prediction. This paper provides various results of using these classification techniques. But the target of each reach work is to achieve better accuracy and better performance in Heart Disease.

REFERENCES

[1]. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66. | [2] Prodromidis, Andreas, Philip Chan, and Salvatore Stolfo. "Meta-learning in distributed data mining systems: Issues and approaches." Advances in distributed and parallel knowledge discovery 3 (2000). | [3] Robnik-Šikonja, Marko. "Improving random forests." Machine Learning: ECML 2004. Springer Berlin Heidelberg, 2004. 359-370. | [4] Karaolis, M.A. Moutiris, J.A. Hadjipanayi, D. Pattichis, C.S., Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees‖ IEEE Transactions on Information Technology in Biomedicine, pp 559 -566, May 2010. | [5] Mrs. G.Subbalakshmi, Mr.M.Chinna Rao "Decision Support in heart disease prediction system using naïve bays",IJCSE Indian journal of computer sciense and engineering, ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011. | [6] S. Vijiyarani et. al., "An Efficient Classification Tree Technique for Heart Disease Prediction" ,International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications (IJCA) (0975 – 8887), 2013 | [7] Deepali Chandna "Diagnosis of Heart Disease Using Data Mining Algorithm", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1678-1680. | [7] E. E. Tripoliti, D. I. Fotiadis, and G. Manis, "Dynamic construction of random forests: Evaluation using biomedical engineering problems,", presented at the 10th Int. Conf. Inf. Technol. Appl. Biomed. Corfu, Greece, 2010. | [8] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, pp. 1–3, 1950. | [9] Shanta kumar, Patil, Y.S.Kumara swamy, "Predictive data mining for medical diagnosis of heart disease prediction" IJCSE Vol .17, 2011 | [10] S. Bernard, L. Heutte, and S. Adam, "On the selection of decision trees in random forests," in Proc. IEEE-ENNS Int. Joint Conf. Neural Netw., 2009, pp. 302–307. | [11] Latinne, Patrice, Olivier Debeir, and Christine Decaestecker. "Limiting the number of trees in random forests." Multiple Classifier Systems. Springer Berlin Heidelberg, 2001. 178-187. |