Prediction and Diagnosis of Heart Disease by Data Mining Techniques

Boshra Bahrami, Mirsaeid Hosseini Shirvani*
Department of Computer Engineering, Sari Branch, Islamic Azad University
Sari, Iran
Boshrabahrami_znu@yahoo.com; *corresponding author Email: Mirsaeid_hosseini@yahoo.com

***Abstract*—**Today, Data mining is rapidly growing in wide range of applications. One of the important data mining fields is medical data mining. There is a wealth of data available in healthcare but there is no effective analysis tool to discover hidden relationships in data. Although millions of people die of heart disease annually, application of data mining techniques in heart disease diagnosis seems to be essential. Discovered Knowledge can help physicians in diagnosis of heart disease. The objective of this paper is to evaluate different classification techniques in heart disease diagnosis. Classifiers like J48 Decision Tree, K Nearest Neighbors(KNN), Naive Bayes(NB), and SMO are used to classify dataset. After classification, some performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. The comparison results show that J48 Decision tree is the best classifier for heart disease diagnosis on the existing dataset.

Keywords—Data Mining; Classification; Heart Disease; Decision Tree

I. Introduction

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods [1].Knowledge discovery and data mining are concepts that are applied in business more than a decade. By development of data mining technology, it is not only extensively applied in commercial purposes, but also successfully applied in many different applications like medical tasks, for examples in intensive care medicine analysis, time dependency patterns mining in clinical pathways, breast cancer screening and diagnosis of heart disease [2]. Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data[3]. According to the World Health Organization, 12 million deaths are caused by heart diseases and stroke in the world annually, 50 percent of which can be prevented by controlling risk factors. Heart diseases are expected to be the main reason for 35 to 60 percent of total deaths expected worldwide by 2025. In Iran 44 percent of death is because of heart disease and it is the second reason of death in country.

This paper is organized as follows: section 2 is literature review, section 3 is about related works, section 4 describes the dataset and proposed method, in section 5, experimental results and performance of proposed method on dataset is shown and in section 6 final conclusion from evaluation and analysis and comparison if different data mining techniques are explained.

II. Literature review

Data mining is the process of extracting hidden knowledge from data. It can reveal the patterns and relationships among large amount of data in a single or several datasets[4]. In other words Data mining is one of the steps of knowledge discovery for extracting implicit patterns from vast, incomplete and noisy data [5]. It is a field with the confluences of various disciplines that has brought statistical analysis, machine learning techniques, artificial intelligence and database management systems together to address the issues [6].

Classification and clustering have been two main issues in the data mining tasks. Classification is the supervised learning task of finding the common properties among a set of objects in a database and classifying them into different classes [7]. Classification is closely related to clustering, since both put similar objects into the same category. In classification, the label of each class is a discrete and known category, while the label is an unknown category in clustering problems [8]. Clustering was thought as unsupervised classification [9]. Since there are no existing class labels, the clustering process summarizes data patterns from the dataset. Usually medical data mining has been treated as a classification problem, which is to search for optimal classifier to classify patient and healthy. Today researchers are using data mining techniques in the diagnosis of several diseases such as diabetes, stroke, cancer, and heart disease. There are many different classification techniques that some of them are discussed in following.

A. J48 DecisionTree Classifier

Decision tree is a kind of classifying and predicting data mining technology, belonging to inductive learning and supervised knowledge mining technology. As decision tree is advantageous in fast construction and generating easy-to-interpret If-Then decision rule, it has become the most widely applied technique among numerous classification methods[10]. Decision tree algorithm has been applied in many medical tasks, for examples, in increasing quality of dermatologic diagnosis [11], predicting essential hypertension [12], and predicting cardiovascular disease [13]. Decision tree is one of the most popular tools for classification and prediction. Production of a decision tree is an efficient method for classification of data. This tree using a top-down strategy to build a test on each node. J48 decision tree method is the implementation of c4.5 decision tree in weka data mining tool. J48 decision tree supports continuous and discrete features. It can also manage features with missing value.

B. *KNN Classifier*

KNN classification is based on the closest training examples in the feature space [14]. This is a type of instance based learning technique also called non parametric lazy algorithm. This technique does not use any assumptions on the data distribution, and hence, is called non parametric. In this algorithm, the k-nearest neighbors are estimated and majority voting is performed. During this, the class which is found most common among the k neighbors is assigned as the class for the new data. K Nearest neighbor is one of the most popular classification techniques introduced by Hodges and fix [15]. Without any additional data, classification rules are generated by the training samples themselves.

C. *SMO Classifier*

Support vector machine is a class of machine learning algorithms that can perform pattern recognition and regression based on the theory of statistical learning and the principle of structural risk minimization [16]. Implementation of support vector machine in weka data mining tools is Sequential Minimal Optimization, which is an algorithm for efficiently solving the optimization problem that arises during the training of Support Vector Machines. It was introduced by John Platt in 1998 at Microsoft Research. SMO is widely used for training SVM. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers [17].

D. *Naive Bayes Classifier*

Naïve Bayes is a data mining technique that shows success in classification of diagnosing heart disease patients [18]. Naïve Bayes is based on probability theory to find the most likely possible classifications [19]. According to Bayesian theorem, the probability of a set of data x_t belonging to c is:

$$P(C|X_t) = \frac{p(C)p(X_t|C)}{p(X_t)} \tag{1}$$

Based on (1), Bayesian classifier calculates conditional probability of an instance belonging to each class, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. In knowledge expression, it has the excellent interpretability same as decision tree, and is able to use previous data to build analysis model for future prediction or classification [20]. The Bayesian classifier has been applied in many medical issues, for examples, in measuring quality of care in psychiatric emergencies [21] and assisting diagnosis of breast cancer [22].

III. Related works

In the diagnosis of heart disease, large number of works carried out. Researchers have been investigating the use of data mining techniques to help health care professionals.

Yeh et al. [23] have investigated the problem of heart disease diagnosis by data mining. In this study authors acquired 493 valid samples from this cerebrovascular disease prevention and treatment program, and adopted three classification algorithms, decision tree, Bayesian classifier and back propagation neural network, to construct classification models, respectively. After analyzing and comparing, classification efficiencies such as sensitivity and accuracy, the decision tree constructed model was chosen as the optimum predictive model for heart disease.

Alizadehsani et al. [24] have studied the diagnosis of coronary artery disease. They used a dataset by 303 samples and enriched the dataset by applying a feature creation method. Then Information Gain and confidence were used to determine the effectiveness of features on CAD. In this study, several algorithms including Naive Bayes, SMO, Bagging, and Neural Network were applied on dataset. Results showed that SMO along with the feature selection and feature creation method achieved the best accuracy.

Giri et al. [25] proposed a methodology for the automatic detection of normal and Coronary Artery Disease conditions using heart rate signals. PCA, LDA and ICA were applied on the dataset in order to reduce the data dimension. The selected sets of features were fed into four different classifiers: SVM, GMM, PNN and KNN. Results showed that the ICA coupled with GMM classifier combination resulted in highest accuracy, sensitivity and specificity, compared to other data reduction techniques and classifiers.

Akhil jabbar et al. [26] proposed new algorithm which combines KNN with genetic algorithm for effective classification. Researchers used genetic search as a goodness measure to rank the attributes. Classification algorithm is built based on evaluated attributes. Experimental results showed that their algorithm enhanced the accuracy in diagnosis of heart disease.

IV. Design of predictive model

In this approach at first the existing standard dataset of heart disease pre-processed and prepared. At this stage irrelevant features that may have a negative effect on performance of predictions or may increase complexity of calculations are removed from dataset and the essential filters are applied on data. Then according to the remaining important features, different models designed for heart disease diagnosis. At this point main classification operation begins and classification algorithms such as, decision tree,k-nearest neighbors, naive Bayes and SMO are used to classify existing dataset. After classification, some of the important performance measures calculated for each method. Then after evaluation, analysis and comparison, the best classification technique for existing heart disease is introduced. Fig.1 shows the sequential overview of proposed approach. Steps 3-5 repeat for each classification method.

A. *Dataset Description*

This dataset contains 209 records and 8 features that is collected from a hospital in Iran, under control of health ministry. It is a standard dataset and so far no data mining operation is done on this dataset. Description of dataset features is given in Table 1. Data is from one resource so there is no need of integration operations. Also all the features in all the 209 samples contain value and there is no missing value problem.

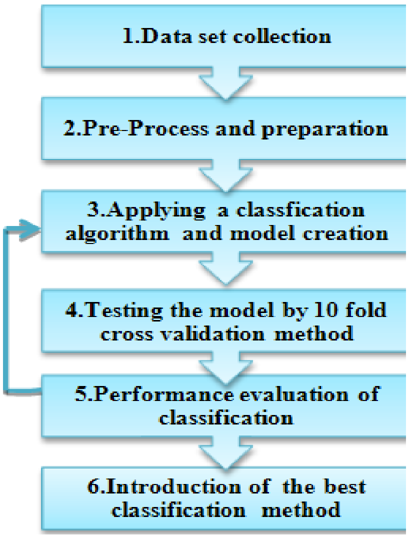
1.Data set collection
2.Pre-Process and preparation
3.Applying a classfication algorithm and model creation
4.Testing the model by 10 fold cross validation method
5.Performance evaluation of classification
6.Introduction of the best classification method

Fig.1.*Sequential overview of proposed approach*

TABLE1.DATASET DESCRIPTION

#	NAME	POSSIBLE VALUES
1	Age	NUMERIC
2	Chest_pain_type	ASYMPT, ATYP_ANGINA, NON_ANGINAL,TYP_ANGINA
3	rest_bpress	NUMERIC
4	blood_sugar	TRUE, FALSE
5	rest_electro	Normal,left_vent_hyper, st_t_wave_abnormality,
6	max_heart_rate	NUMERIC
7	exercice_angina	YES, NO
8	Disease	NEGATIVE, POSITIVE

B. *Feature Selection*

There are so many different methods for feature extraction, In this research Gain Ratio Attribute Evaluation method and ranker search is used as feature selection method. According to the dataset four insignificant features removed from the dataset. finally, four features Chest_pain_type, max_heart_rate, exercice_angina and target feature, Disease, selected as desired features.

C. *Classification Operation*

In this phase, data is ready for applying classification algorithm. After model creation from training data, classification operation is performed on test data. Then some of the most important performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. This study employed 10-fold cross-validation in classification model construction and efficiency evaluation. This method increases the validation of classification and prevents from random and invalid results.

V. ANALYSIS AND EVALUATION

In this study, WEKA that is a powerfull data mining tool, was used to apply the data mining algorithms. In this section experimental results from implementation of selected classification algorithms, j48 decision tree, Naive Bayes, KNN and SMO on heart disease dataset are analyzed and compared. After comparison, results showed that the best classification accuracy is 83.73% that achieved by j48 decision tree. Table 2 shows classification efficiency indicator values of models constructed with four algorithms. Comparison of accuracy that is achieved by four classifiers, is shown in Fig.2. Although accuracy is the most common measure in classification performance, other important performance measures such as sensitivity, Specificity, F-Measure, precision and ROC indicators considered to evaluate and compare classification efficiency of four selected algorithms. Fig.3, shows the Comparison of sensitivity, specificity, precision, F-measure and area under Roc curve, achieved by four classification algorithms on existing heart disease dataset.

TABLE2. PERFORMANCE OF THE CLASSIFIERS

EVALUATION CRITERIA	CLASSIFIERS			
	J48	KNN	NB	SMO
Correctly classified instances	175	173	171	173
Incorrectly classified instances	34	36	38	36
Accuracy	83.732	82.775	81.818	82.775

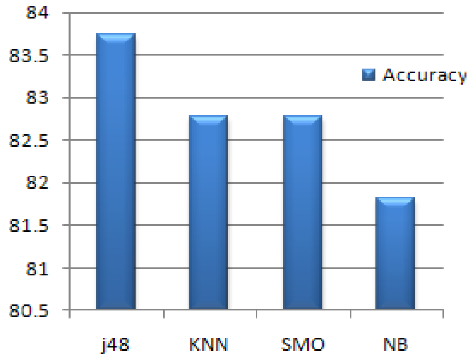

Fig.2.*comparison of accuracy achieved by four classifiers*

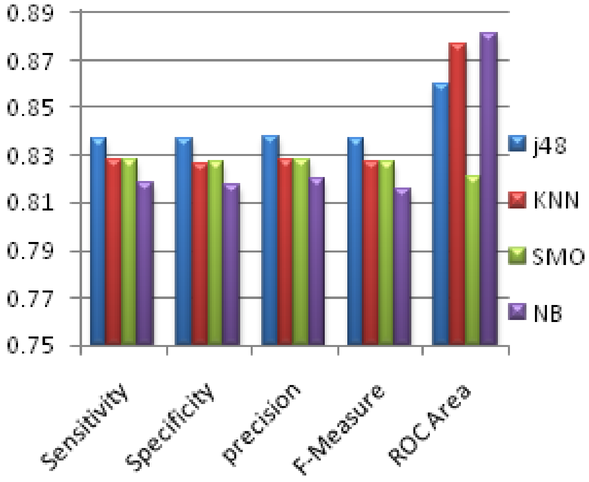

Fig.3.*comparison of other performance measures in classifiers*

VI. CONCLUSION

In this study four different classification algorithms applied on existing heart disease dataset. The Gain Ratio evaluation technique used as feature selection technique and four features extracted from dataset. Then preprocessed datasets, used to test the four classifiers using 10-folds cross validation. Six different performance measures considered for classifiers. Results of comparison showed that j48 decision tree achieved the highest value in accuracy, sensitivity, specificity, F-measure and precision performance measures. The optimum heart disease predictive model obtained in this study, adopts j48 decision tree as classification algorithm. Hence, it is a suitable candidate for testing in a clinical environment and implementing in decision support systems for helping physicians and healthcare professionals in diagnosis of heart disease.

REFERENCES

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

[2] Ganzert, S. and Guttmann, J, " Analysis of respiratory pressure–volume curves in intensive care medicine using inductive machine learning,"Artificial Intelligence in Medicine, 2002.

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

[4] Witten, Ian H., and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques,"Morgan Kaufmann,2005.

[5] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P."From data mining to knowledge discovery in databases,"Artificial Intelligence Magazine, 1996.

[6] Venkatadri, M., and Lokanatha, C. R. "A review on data mining from past to the future,"International Journal of Computer Applications, 19-22,2011.

[7] Chen, M. S., Han, J., and Yu, P. S." Data mining: An overview from a database perspective," IEEE Transactions on Knowledge and Data Engineering,8, 866–883,1996.

[8] Xu, R. and Wunsch, D. "Survey of clustering algorithms,"IEEE Transactions on Neural Networks, 2005.

[9] Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323,1999.

[10] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. "Discovering data mining: From concept to implementation,". New Jersey: Prentice Hall,1997.

[11] Chang, C.-L., and ChenC.-H."Applying decision tree and neural network to increase quality of dermatologic diagnosis,"Expert Systems with Applications, 4035-4041.,2009.

[12] Ture, M., Kurt, I., Kurum, A. T., and Ozdamar, K,"Comparing classification techniques for predicting essential hypertension,".Expert Systems with Applications, 2005.

[13] Eom, J.-H., Kim, S.-C., and Zhang, B.-T. "AptaCDSS-E: A classifier ensemblebased clinical decision support system for cardiovascular disease level prediction," Expert Systems with Applications,2008.

[14] Acharya, U. R., Sree, S. V.,. Chattopadhyay, S, W. Yu and A.P.C. Alvin, "Application of recurrence quantification analysis for the automatic identification of epileptic EEG signals,"International Journal of Neural Systems,2011.

[15] Lee, I.-N., S.-C. Liao, and M. Embrechts, "Data mining techniques applied to medical information," Med. Inform, 81-102,2000.

[16] Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., and Balaji, P. V."A support vector machine-based method for predicting the propensity of a protein to be soluble or to form

inclusion body on overexpression in escherichia coli,"Bioinformatics, 2006.

[17] Platt, John, "Sequential minimal optimization: a fast algorithm for training support vector machines,"Technical Report Microsoft Research, 1998.

[18] Sitar-Taut, V.A., et al., "Using machine learning algorithms in cardiovascular disease risk evaluation,"Journal of Applied Computer Science & Mathematics, 2009.

[19] Yadav, S. K., and Pal, S."Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification,"World of Computer Science and Information Technology (WCSIT), 2012.

[20] Loether, H. J., and McTavish, D. G." Descriptive and inferential statistics: An introduction (4th ed.),"Needham Heights, MA: Allyn and Bacon,1993.

[21] Gustafson, D. H., Sainfort, F., Johnson, S. W., and Sateia, M, "Measuring quality of care in psychiatric emergencies: Construction and evaluation of a Bayesian index," Health Services Research, 1993.

[22] Wang, X.-H., Zheng, B., Good, W. F., King, J. L., and Chang, Y.-H. "Computer assisted diagnosis of breast cancer using a data-driven Bayesian belief network,"International Journal of Medical Informatics, 1999.

[23] Yeh, D.-Y., Cheng, C.-H., and Chen, Y.-W."A predictive model for cerebrovascular disease using data mining," Expert Systems with Applications, 2011.

[24] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, "A data mining approach for diagnosis of coronary artery disease,"Computer Methods and Programs in Biomedicine, 2013.

[25] Giri, D.and Acharya, U. R,. "Automated diagnosis of Coronary Artery Disease affected patients usingLDA, PCA, ICA and Discrete Wavelet Transform." Knowledge-Based Systems, 2013.

[26] Jabbar,M.A,Deekshatulu,B.LandChandra,Priti ,"Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,"Procedia Technology, 85– 94,2013.