IJE TRANSACTIONS A: Basics Vol. 29, No. 7, (July 2016) 921-930

International Journal of Engineering

Journal Homepage: www.ije.ir

Analysis of Pre-processing and Post-processing Methods and Using Data Mining to Diagnose Heart Diseases

H. Hamidi*, A. Daraei

Department of Industrial Engineering, Information Technology Group, K. N.Toosi University of Technology, Tehran, Iran

PAPER INFO

Paper history:
Received 27 March 2016
Received in revised form 30 April 2016
Accepted 02 June 2016

Keywords:
Data Mining
Heart Disease
Diagnosis
Prognosis
Treatment

ABSTRACT

Today, a great deal of data is generated in the medical field. Acquiring useful knowledge from this raw data requires data processing and detection of meaningful patterns and this objective can be achieved through data mining. Using data mining to diagnose and prognose heart diseases has become one of the areas of interest for researchers in recent years. In this study, the literature on the application of classification algorithms for heart disease will be reviewed. The present study is an attempt to evaluate the studies carried out in this field so that the results of this review may lead to development of a clear view on the future studies. Here, first, the major medical tasks are specified and then, each article is investigated based on these tasks. Finally, some results, in terms of frequency algorithms in the use of classification algorithms, pre-processing and post-processing methods, will be provided. In this study, 49 articles obtained from similar studies with related subject matters, (from 2003 to 2015) are collected and reviewed. Obviously, the number of articles on applications of classification algorithms in heart disease is quite significant, therefore, it is impossible to review all of them in the present study. It is hoped that this study can provide results that pave the path for future research and further developments in this area.

***doi**: 10.5829/idosi.ije.2016.29.07a.06*

1. INTRODUCTION

The increasing growth of medical data in recent years, has generated a huge amount of data in the medical field and this could be regarded as the main reason for growing inclination towards the use of data mining in this area. Using data mining techniques allows for transformation of meaningless data into useful information and knowledge [1]. Data mining can be regarded as a tool for acquiring knowledge from raw and meaningless data in the medical field, and in the field of heart diseases in particular. According to the World Health Organization (WHO), heart disease is the leading cause of death in the world and accounts for 31% of deaths in the world. WHO estimates that the rate of death from heart disease in 2030 will reach 23.6 million[1] . According to statistics, the use of data mining methods to achieve useful patterns is necessary. In recent decades, studies on the application of data mining, and particularly classification methods as the most widely used data mining method [1] for heart disease, has been growing rapidly. The aim of this study is to review the applications of data mining classification algorithms for heart disease. Although many review studies have been carried out in the field of data mining applications in the medical field, the number of review studies carried out on data mining applications in heart diseases is very limited. Due to small number of such studies, in the present study attempts are made to review previous studies on the applications of classification algorithms for heart diseases. For this 49 extracted articles, published from 2003 to 2015, were studied. This study is organized as follows: in the second section, the basic concepts and

*Corresponding Author's Email: *hamidi@kntu.ac.ir* (H.Hamidi)

1. World Health Organization, "WHO | Cardiovascular diseases (CVDs)".2016.[Online].Available:http://www.who.int/mediacentre/factsheets/fs317/en/.

Please cite this article as: H. Hamidi, A. Daraei, Analysis of Pre-processing and Post-processing Methods and Using Data Mining to Diagnose Heart Diseases, International Journal of Engineering (IJE), TRANSACTIONS A: Basics Vol. 29, No. 7, (July 2016) 921-930

definitions related to the topic are discussed. In the third part, the research methodology is described. The fourth part deals with the review of literature. Discussion of the literature is provided in part five and finally, the sixth part of the study deals with the conclusion.

2. BASIC CONCEPTS AND DEFINITIONS

2. 1. Heart Disease Heart disease was known as the leading cause of death among other non-communicable diseases in 2012. In 2012, this disease accounted for 46% of deaths from non-communicable diseases. In the other words, 31% of deaths in the world are caused by heart diseases[2] . Different types of heart disease include: Coronary Artery Disease, Angina Pectoris, Congestive Heart Failure, Cardiomyopathy, Congestive Heart Disease, Arrhythmia and Myocarditis [2].

2. 2. Data Mining Data mining is a part of knowledge discovery in databases (KDD) [3]; in the other words, data mining is considered as a KDD process with three stages: Data Pre-processing, Data Modeling and Data Post-Processing [4].

2. 2. 1. Data Preprocessing At this stage, the raw data are prepared for knowledge discovery processing [6]. In practice, 60 to 90% of the data mining time is spent on data preparation in the Data Pre-processing stage [5]. Pre-processing methods commonly used in heart disease are as follows:

- **Data Cleaning** Noise in the data can lead to errors in the data mining process, and these noises should be removed. The outlier data, which differs from the problem's data, can also affect the outcome and should be removed. In addition, missing value, which is defined as the empty cells in the data can be deleted or ignored or be replaced with proper estimate [6-15].

- **Attribute Transformation** Data can have different moods and forms that should transform into suitable forms for data mining [9, 14-18].

- **Discretization and Binarization** This is the conversion of continuous data to discrete and binary data, so they can be used in the data mining process [8, 19].

- **Dimensionality Reduction** Many data and features are involved in the medical field. Reduction of these data and features can lead to better model results and higher accuracy. This method removes the irrelevant or redundant features while keeping and using the best features to avoid curse of dimensionality and irrelevant features [20, 21]. Feature selection is one of the most important and most commonly used methods of dimensionality reduction in literature [13-18, 22-40].

2. 2. 2. Data Modeling Data modeling tasks in the data mining process are divided into two categories: predictive tasks and descriptive tasks. The algorithms of the predictive category include classification algorithms for discrete data, and regression algorithms, for continuous data [41], which are learned through supervised learning process [1]. The descriptive categories include clustering algorithms and association rule mining algorithms [41], which are learned through non-supervised learning processes. The classification methods in the predictive category include algorithms which determine the class labels in accordance with the training data with specific labels. In classification, data are divided into two sections and the modeling process takes place based on these sections. The first section includes the training data for learning, and the second section includes the test data used to validate the model [1, 42].

2. 2. 3. Data Post-processing At this stage, different evaluation methods are used to assess the model's efficiency. These methods generally include scalar methods such as accuracy, sensitivity and specificity, and graphical methods such as the ROC curve. The advantage of scalar methods is their easy use; however, these methods do not consider all aspects for evaluation. Despite their difficult implementation, the graphic methods offer better results for evaluation [43]. The best data mining purposes in the medical field include reduction of human errors, improvement of efficiency and accuracy, reduction of time, reduction of costs and helping doctors make medical decisions [1].

2. 3. Medical Data UCI Data Repository datasets[3] are used in most of the articles. These datasets are similar and include 13 features for diagnosis of heart diseases. These features include age, gender, type of chest pain, blood pressure while resting, serum cholesterol, maximum heart rate during exercise, fasting blood sugar, ECG results, ST elevation, ST depression, the number of vessels specified by X-ray, and the type of induced pain or defect.

Cleveland is the most famous dataset from UCI Data Repository in heart diseases. This dataset includes 303 features and 76 features, 13 of which are used in the above-mentioned articles.

2. World Health Organization, "Global status report on noncommunicable diseases 2014", 2016. [Online]. Available: http://www.who.int/nmh/publications/ncd-status-report-2014/en/.

3. Archive.ics.uci.edu, "UCI Machine Learning Repository", 2016. [Online]. Available: http: //archive.ics.uci.edu/ml.

3. RESEARCH METHODOLOGY

In this study, 49 articles published from 2003 to 2015 were reviewed. To find these articles, leading publishers such as Elsevier, IEEExplore, Springer and PubMed and other magazines and search engines were searched. These articles were reviewed through extensive search, advanced search, concept extraction and analysis of the literature. In the extensive search, all articles in the field of data mining and heart disease are searched. In the advanced search, the titles with the exact classification terms and heart disease terms, and then abstract and introduction of the relevant articles are searched and the articles that have only partially dealt with them are removed. Finally, 49 extracted articles are analyzed and reviewed. As mentioned above, 49 articles regarded as representative articles in the area of classification applications for heart diseases are investigated and reviewed in this study. So, it can't be claimed that all the relevant articles in this field have been covered in this review. In this study, attempts are made to review the majority of the articles, however this can be considered a limitation for the present study.

4. REVIEW OF LITERATURE

The most important tasks in the field of medicine and, consequently, heart disease, can be divided into three categories: Diagnosis, Prognosis and Treatment [44].
In this section, articles are studied in terms of medical tasks and the studies related to each medical task are briefly discussed.

4. 1. Diagnosis Diagnosis refers to detection of the disease nature using tests and symptoms that could lead to treatment of the patients [41].

Kumar and Sahoo [3] proposed a new algorithm including NB and genetic algorithm to improve the classification of cardiovascular diseases. In the evaluation phase, the accuracy of this method for Cleveland dataset was 100%. Pedreira et al. [6] used the Artificial Neural Network (ANN) to diagnose heart disease. In that study, instead of using the Euclidean distance criterion, a weighing-based criterion is used in the K Nearest Neighbors (K-NN) algorithm. The dataset used in this study was obtained from the Data Repository UCI with accuracy of 80%. Arif [10] used Back Propagation Neural Network to diagnose Myocardial Infarction (MI) and determine its location. The main indices in that study included Q and T waves as well as ST level Elevation or Depression. Principal component analysis (PCA) was used to extract the features. The classification method was also separately used to diagnose MI and determine its location. Sensitivity and specificity were used to assess the diagnosis of MI. Shao et al. [13] used logistic regression, MARS, artificial neural network and the rough sets and proposed LR-ANN, MARS-ANN and RS-ANN hybrid models for classification. Neural Network along with each of these three methods, was applied on heart disease datasets including 899 patient records. The logistic regression method, multivariate adaptive regression splines method, and Rough set each selected 12, 6 and 10 important variables from this dataset. ANN for each of these scenarios is used for classification. After the evaluation, the diagnosis accuracy rate for the LR-ANN, MARS-ANN, and RS-ANN proposed methods was equal to 78.57, 82.14 and 79.5%, respectively. A system which uses data mining methods and imperialist competitive algorithms (ICA) was proposed [14]. The accuracy of this method was equal to 94.92%. A bagging algorithm was used [17] to determine the risk factors in heart patients and also to compare the effectiveness of the bagging methods compared to the decision tree. The experiments were conducted in two phases: in the first phase, a J48 decision tree algorithm with 10-fold cross validation was used. In the second phase, the bagging algorithm and the J48 decision tree were used together. After the evaluation, it was found that the bagging method was more effective than the decision tree approach. Fei [22] used SVM model and particle swarm optimization to diagnose cardiac arrhythmia. The accuracy of that method was 95.625% and was obtained by PSO-SVM, using optimized parameters.

Dennis and Muthukrishnan [19] proposed a genetic fuzzy system; this model is a fuzzy system that has been upgraded by a genetic algorithm-based learning process. Highest accuracy rate for heart disease after 20 repetitions was equal to 76.67 and 59.9%, respectively. An abnormality detection system was proposed [20] which detects abnormality in the heart and its walls based on real datasets. After the feature selection, SVM classification was applied to the data. The evaluation results show that the selection of 3 main features leads to achievement of high efficiency. Esmaileyan and Marvi [21] proposed a heart disease diagnosis using a new K-NN-based weighing method that is used as a pre-processing stage, before the classification. The method used in that study was an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism. The best accuracy obtained there was k =15 and sensitivity and specificity levels equal to 92.3 and 78.57%, respectively. Conforti [25] used a Support Vector Machine (SVM) and 5 linear, Gaussian, Laplacian, polynomial and sigmoid core functions, to categorize the patterns for detection of acute myocardial infarction. On average, the best accuracy in 7 different feature selection modes was equal to 85.85%, for linear and polynomial function, 82.64% for the Gaussian, 86.43% for Laplace and 82.4% for the sigmoid. Naïve Bayes algorithms, Decision Tree, K-NN as well as 92 features were used [26] to detect Coronary Artery

Diseases. 3000 sample datasets were used in that study and the comparison results were based on 10-fold cross validation. Finally, the obtained accuracy for NB formula, decision tree and KNN was equal to 52.33, 52 and 45.67%, respectively. According to the results, the NB formula gave the highest accuracy. Babaoglu et al. [29], the SVM algorithm was used to diagnose heart diseases. In that study, a genetic algorithm was also used to select a subset of features. The accuracy obtained for that method was 81.46%. Peter and Somasundaram [32] proposed a fuzzy rule-based system which served as a decision support system to detect Coronary artery diseases. The classification failure rate in case no feature was removed, and after feature selection processes was equal to 17.07 and 17.04%, respectively. The accuracy of that method, before and after applying the group classification strategy was equal to 81.85 and 84.44%, respectively. A hybrid method was used [34] to detect Coronary artery diseases. In that hybrid system, a combination of particle swarm optimization (PSO) and the proposed method was used. After using PSO to optimize the fuzzy membership function, accuracy of 93.27% was achieved. A way for mining meaningful information from the heart disease datasets for better detection of heart disease was proposed by Santhanam and Ephzibah [35]. In that study, the most important features in the dataset were selected using principal component analysis and regression techniques. These features are used for classification and prediction by regression and feed-forward neural network classifiers. Accuracy obtained using principal component analysis for regression analysis, and for the feed-forward neural network was 88.5 and 90.54%, respectively.

Subanya and Rajalaxmi [37] proposed a meta-heuristic algorithm to determine an optimal subset of features with increased accuracy of classification in diagnosis of cardiovascular diseases. In this study, the bee colony was used to optimize the feature selection processes. The accuracy of that method with 7 selected features by ABC, was equal to 86.76%. Nguyen et al. [40] proposed an automatic classification method for medical data which takes advantage of Wavelet Transformation and Fuzzy Logic Type 2. The Cleveland dataset was used in that study; 6 features were removed due to incomplete information. The best accuracy was reached in case of 3 selected features by the wavelets. The results showed that, increased number of selected features by the wavelet led to reduction of the model's accuracy. Yan et al. [45] offered a computational model based on Multi-Layer Perceptron Neural Network to be used in the decision support system. In the evaluation stage, chronic heart disease had the highest accuracy in the classification and CAD had the lowest accuracy. Das et al. [46] proposed a neural network ensemble model to detect heart diseases. Neural network was used to classify the data. The accuracy of the proposed model was equal to 89.01% which is higher compared to the previous models. Chu et al. used the Bayes model to diagnose CAD [47]. Pal et al. [48] attempted to detect CAD in the early stages. Fuzzy data mining method was used in that study. Methodology of that study included fuzzy linguistic labels, fuzzy rule databases and an organizational unit of defined rules. Finally, after efficiency evaluations, that method reached the accuracy and sensitivity of 84.20 and 95.85%, respectively.

Sanz et al. [49] developed a classifier to determine the risk of cardiovascular disease in the next 10 years. In that study, a data set of 828 cases was considered. Finally the proposed method, compared to classic fuzzy classifiers improved the percentage of patients who had been diagnosed correctly by 3%. A hybrid learning algorithm for MLP fuzzy neural network was proposed [50] that used a combination of two Gravity Search meta-heuristic algorithms and PSO. The accuracy of the proposed method for heart disease was 76.97%. Kumar [51] investigated the performance of soft computing techniques to classify cardiac arrhythmia. Decision tree and CART clustering technique, MLP and SVM for classification were applied on the datasets. Finally, the obtained accuracy of the CART, SVM with RBF kernel function, SVM with polynomial function, and MLP was equal to 88.23, 92.16, 62.9 and 88.23%, respectively. SVM with RBF kernel function had the best performance. Safdarian et al. [52] attempted to detect and locate MI. The classification methods used in that study included probabilistic neural network (PNN), K-NN, MLP, and Naïve Bayes (NB). The dataset used in that study included 549. Finally, the highest accuracy for the diagnosis of MI was obtained by the NB method. The obtained accuracy was lower for the diagnosis of MI and its location and the best accuracy (76.67%) was achieved by PNN. Abuhasel et al. [53] proposed a hybrid approach using Adaboost Ensemble and Neural Network with Fuzzy Membership Function (NEWFM) for classification of medical data. In the first phase in the proposed algorithm, the training data's weight with similar values were initialized. Then, that process was repeated according to the number of classifiers. The accuracy of the heart disease was equal to 97.4%. In another work [54] Standard Additive Model (SAM) was merged with the genetic algorithm. That model is called GSAM. Wavelet Transformation was used to extract discriminative features for high-dimensional data sets. The data set used for the heart disease in that study was the Cleveland dataset. The accuracy of the proposed method was equal to 78.78%.

4. 2. Prognosis Prognosis refers to prediction of disease outcomes (whether the person will get the disease or not) [44]. This factor can also predict the survival or death of individuals suffering from the disease [1].

Various data mining models was used by Srinivas et al. [8] to prognose heart diseases. They obtained their data from the UCI repository. They used four classification methods such as rule-based classification, decision trees, NB and artificial neural network in that study. The results showed that the accuracy of NB heart disease prognosis was equal to 84.14%. A method for CVD disease prognosis was proposed [9]. The proposed method was the neural network for heart disease prognosis which made use of genetic algorithm to achieve optimal weights. After the neural network weight optimization using genetic algorithms, the evaluation results showed the accuracy of 94.17% for the model. In another work [11] unlike previous studies, not only the features were not reduced, but a larger number of features were used to prognose the heart disease. Three data mining methods including decision trees, NB and artificial neural network were used to classify and analyze the dataset of heart diseases. The results show that in NB, fewer number of features led to higher accuracy, but in the neural network and decision tree, greater number of features brought about better results.

Anooj [12] attempted to prognose the risk of heart disease in patients. After classification of evaluation results, the highest accuracy was obtained using Cleveland datasets. Bashir [16] used the Ensemble method including NB, linear regression, quadratic discriminant analysis, instance based learner and SVM to prognose heart diseases. Five data sets used in that study were obtained from publicly accessible databases, and were used for classification after the feature selection. The results showed the superiority of that method to other conventional classification techniques. Bhatla [18] compared the different classification techniques, including neural networks, decision trees and NB applied on heart diseases data and found that the neural network act better than other classification methods. Anbarasi et al. [24] tried to provide a more accurate prognosis of heart diseases incidents with reduced number of features. After feature selection, J48 decision tree, NB and classification using clustering were used to diagnose heart disease. In this study, the J48 decision tree classifier provided better results compared to the other two models.

The accuracy of NB, decision trees and classification via clustering was 96.5, 99.2 and 88.3% respectively. Conforti et al. [25] proposed a system to prognose the severity of CVD diseases. The proposed method was a combination of fuzzy logic, neural network and genetic algorithm. After evaluation of the system the average accuracy of 88.35% was achieved. Khemphila [30] classified heart diseases in order to prognose diseases through feature reduction. The Cleveland heart disease dataset was used and using information gain, the number of features reduced from 13 to 8. The difference of accuracy between the two states of features, by neural network was equal to 1.1% in the training mode, and equal to 0.82% in the validation state. Peter and Somasundaram [32] proposed a model to prognose heart disease using data mining techniques. They applied decision tree, NB, K-NN and neural network classification methods, first on the primary dataset and then on the generated dataset after dimensionality reduction. The best accuracy for NB algorithm (85.5%) was achieved in reduced dimensionality state using CFS Subset algorithm. In another work, Peter and Somasundaram [33] proposed as hybrid feature selection-based framework to prognose heart diseases. The aim of that study was to use a new method for feature selection composed of CFS and NB, in which four algorithms NB, j48 decision tree, K-NN and neural network algorithms were used after feature selection. The following accuracies were achieved for classification techniques: NB (83%), decision tree (76.66%), K-NN (75.18%), and neural network (78.14%).

Patel et al. [36] provided a model for prognosis of heart diseases using classification methods. After reduction of features, 3 classification algorithms NB algorithm, j48 decision trees and classification using clustering algorithms were used for classification. The decision tree brought about higher accuracy compared to the other two methods. Krishnaraj and Vinothkumar [38] used the combination of multi-layered feed-forward neural network algorithm with genetic algorithm to prognose heart attack. Finally, in the evaluation stage, the propose method obtained accuracy of 88%.

A method for prognosis of diseases was provided [39]. The study was carried out based on the feature selection and using multilayer back-propagation neural network. The K-NN algorithm was used for classification. Finally, the accuracy of modeling in the primary state with all the features was equal to 93% which was higher than the modelling accuracy after features were reduced to 8.

A method for prognosis of heart diseases using decision trees was proposed [55]. The decision tree used in that study was the gain decision tree. After extraction of rules, decision trees used reduced error pruning to extract the division rules. Masethe and Mesethe [56] used J48, Bayesian networks, simple Bayes, CART and REPTREE methods for classification and development of a model to prognose heart attack. According to the evaluation results, J48, NB and CART classification methods achieved higher accuracy and proved to be better than NB techniques and the Bayesian networks.

4. 3. Treatment Treatment refers to choice of therapy based on suitable drugs and therapeutic methods [44]. Although selection of treatment and medications is important in the medical field and the use of classification techniques can lead to significant results

in this field, very few studies have been carried out on the treatment of heart disease using classification methods.

Kim et al. [57] assessed the state of patients in the treatment of heart failure using decision tree. Neural network, NB and decision tree classification methods were used by Shouman et al. [58] to diagnose heart disease and determine the type of treatment. The results showed that hybrid methods led to achievement of more efficient results.

5. DISCUSSION

According to the articles in the review of literature, analysis is performed based on 4 categories:

• **Medical Task** Of the above-mentioned medical tasks, researchers seem to be more interested in diagnosis rather than prognosis and treatment of heart diseases. Due to the high cost of treatment, surgical or diagnostic procedures for heart diseases in hospitals and life-threatening risks of such diseases, it appears that the prognosis task is very important and needs to be taken into closer considerations. As for treatment task for heart diseases, determination of medications and treatment is of great importance, however less attention has been paid to this task. Figure 1 shows the number of articles on each medical task.

Data Preprocessing As mentioned above, Data Pre-processing is the most important step in knowledge discovery and can have a significant impact on the result of data mining. Figure 2 shows the percentage of articles in terms of Data Pre-processing and the lack of it. The importance of this step can also be understood from this figure. According to previous studies, feature selection is the most common Data Pre-processing method. After that, data cleansing and data transformation are the second most widely used methods in this field. Figure 3 shows the number of articles in terms of pre-processing method uses.

• **Data Modeling** The frequency of classification algorithms, regardless of medical tasks, used in the reviewed articles, is shown in Figure 4.

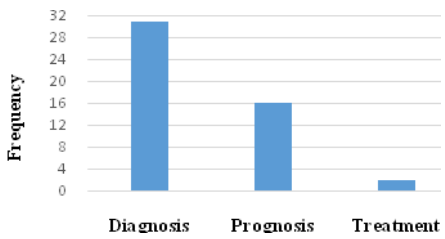

Figure 1. Number of articles on each medical task, in literature

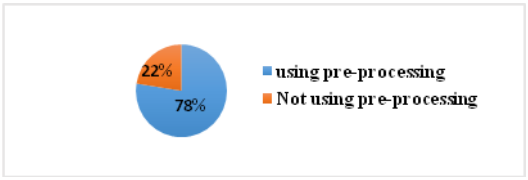

Figure 2. Percentage of articles in terms of using Data Pre-processing, in literature

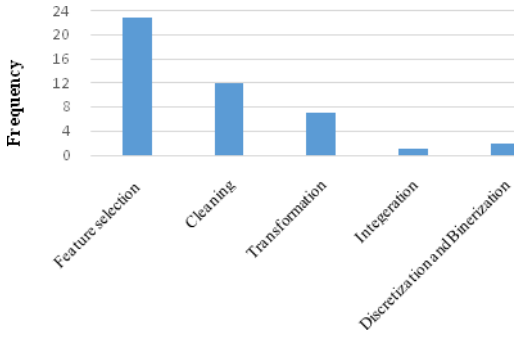

Figure 3. Number of articles in terms of pre-processing method used in literature

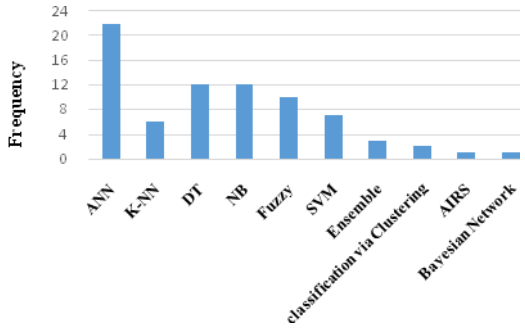

Figure 4. The frequency of classification algorithms, regardless of medical tasks, in literature

In addition, Figure 5 and Figure 6 show the classification algorithms used in the literature in terms of diagnosis and prognosis tasks, respectively. As for the treatment task, due to the small number of studies, no comparison, in the form of charts, is made here. As shown in Figures 4, 5 and 6, neural network, despite lower interpretability, is used in a larger number of articles. Although the decision tree is highly interpretable, its lower use could be due to the fact that in the imbalance data, the tree is biased toward the larger part [1].

On the other hand, it can be seen that the use of decision trees have fallen in recent years, but the use of SVM is on the rise. SVM can lead to good results, but lower applications of this method, compared to the

neural network, can be due to its high susceptibility to noise and missing values. It can be seen in the review of the literature that SVM algorithm is used in articles that have passed the Pre-processing stage [13, 16-18, 20, 22, 25] and [35]. Table 1 lists the advantages and disadvantages of the most important classification algorithms based on the literature. On the other hand, given that the neural network acts like a black box and is not able to show knowledge but enjoys high efficiency, prognosis task is more applicable in diagnosis of heart diseases. Given that in prognosis, the rules can prognose the risk of heart disease, the decision tree is of greater use in prognosis of diseases. Considering the trend of articles in this field, it can be expected that SVM, fuzzy and NB be more extensively used in the future.

• **Data Post-processing** According to Figure 7, it can be figured out that accuracy is used in majority of articles to evaluate the use of classification methods. After accuracy, sensitivity and specificity are the second most commonly used methods in this field.

This could be due to the fact that in the medical field, positive classes and negative classes are not equal in terms of value, and in some cases, depending on the type of problem, sensitivity and specificity can be more important than accuracy.

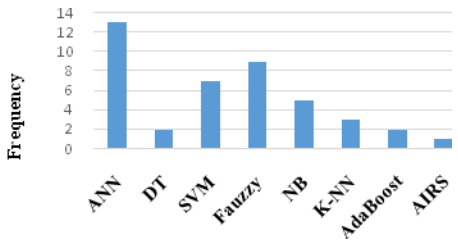

Figure 5. Classification algorithms in terms of diagnosis task, in literature

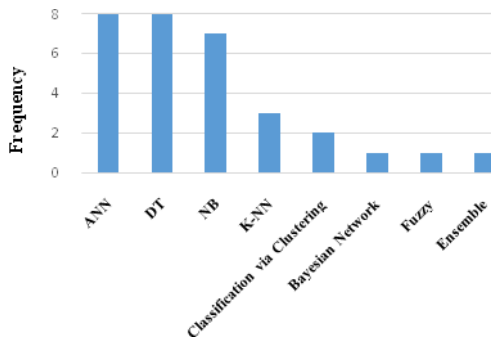

Figure 6. Classification algorithms in terms of prognosis task, in literature

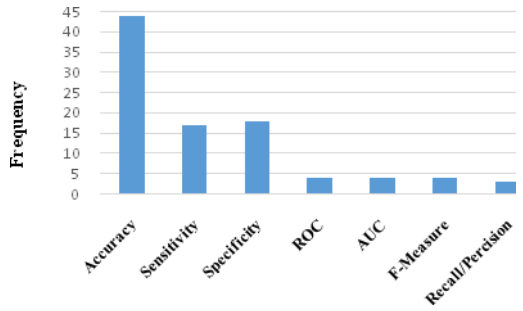

Figure 7. The frequency of evaluation methods in literature

TABLE 1. Advantages and disadvantages of the most important classification algorithms based on the literature

Algorithm	Advantages	Disadvantages
ANN	• High efficiency • Ability to extract complex relations	• High training time • Black-Box
DT	• Appropriate for both numerical and categorical data • High interpretability • Resistant to noise	• Inappropriate for high volume data • Inappropriate for imbalance data
SVM	• Appropriate for high dimensional data • High efficiency	• High training time • High computational cost • Black-Box
NB	• Resistant to noise, missing value and outlier data	• Decrease of accuracy by correlated attributes
K-NN	• High interpretability • Low training time	• Sensitive to noise • Needs high memory • Slow test time

ROC and AUC, despite their high interpretability, have been used less than scalar techniques. Methods that were seen only once in the literature, are not listed in the chart.

6. CONCLUSION

Considering the large volume of raw data that is generated every day in the area of heart disease, the importance of knowledge discovery from this data for use in the medical field can be easily understood. In general, in the medical field, the accuracy and reliability of systems is very important. Despite the high efficiency of classification methods for diagnosis, prognosis and treatment of heart disease, these methods have not reached the necessary confidence level yet. By comparing and analyzing results from previous studies,

this review study can lead to development of a better view on the previous studies, for conduction of future studies.

7. REFERENCES

1. Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E. and Tabar, V. K., "Knowledge discovery in medicine: Current issue and future trend", ***Expert Systems with Applications***, Vol. 41, No. 9, (2014), 4434-4463.
2. Purusothaman, G. and Krishnakumari, P., "A survey of data mining techniques on risk prediction: Heart disease", ***Indian Journal of Science and Technology***, Vol. 8, No. 12, (2015).
3. Kumar, S. and Sahoo, G., Classification of heart disease using naive bayes and genetic algorithm, in Computational intelligence in data mining, Vol. 2, (2015), Springer, 269-282.
4. Fayyad, U. and Uthurusamy, R., "Data mining and knowledge discovery: Making sense out of data", Vol. 39, No. 11, (1996), 24-26.
5. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., "CRISP-DM 1.0 step-by-step data mining guide", (2000).
6. Pedreira, C. E., Macrini, L. and Costa, E. S., "Input and data selection applied to heart disease diagnosis", in Proceedings. IEEE International Joint Conference on Neural Networks, Vol. 4, (2005), 2389-2393.
7. Patil, S. B. and Kumaraswamy, Y., "Intelligent and effective heart attack prediction system using data mining and artificial neural network", ***European Journal of Scientific Research***, Vol. 31, No. 4, (2009), 642-656.
8. Srinivas, K., Rani, B. K. and Govrdhan, A., "Applications of data mining techniques in healthcare and prediction of heart attacks", ***International Journal on Computer Science and Engineering (IJCSE)***, Vol. 2, No. 02, (2010), 250-255.
9. Amma, N. B., "Cardiovascular disease prediction system using genetic algorithm and neural network", in International Conference on Computing, Communication and Applications, IEEE, (2012), 1-5.
10. Arif, M., Malagore, I. A. and Afsar, F. A., "Automatic detection and localization of myocardial infarction using back propagation neural networks", in Bioinformatics and Biomedical Engineering (iCBBE), 4th International Conference on, IEEE, (2010), 1-4.
11. Dangare, C. S. and Apte, S. S., "Improved study of heart disease prediction system using data mining classification techniques", ***International Journal of Computer Applications***, Vol. 47, No. 10, (2012), 44-48.
12. Anooj, P., "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", ***Journal of King Saud University-Computer and Information Sciences***, Vol. 24, No. 1, (2012), 27-40.
13. Shao, Y. E., Hou, C.-D. and Chiu, C.-C., "Hybrid intelligent modeling schemes for heart disease classification", ***Applied Soft Computing***, Vol. 14, (2014), 47-52.
14. Mahmoodabadi, Z. and Tabrizi, S. S., "A new efficient algorithm based on ica for diagnosis of coronary artery disease", ***International Journal of Telemedicine and Clinical Practices***, Vol. 1, No. 2, (2015), 157-173.
15. Bhaskar, N. A., "Performance analysis of support vector machine and neural networks in detection of myocardial infarction", ***Procedia Computer Science***, Vol. 46, (2015), 20-30.
16. Bashir, S., Qamar, U. and Khan, F. H., "A multicriteria weighted vote-based classifier ensemble for heart disease prediction", ***Computational Intelligence***, (2015).
17. Arif, M., Malagore, I. A. and Afsar, F. A., "Detection and localization of myocardial infarction using k-nearest neighbor classifier", ***Journal of Medical Systems***, Vol. 36, No. 1, (2012), 279-289.
18. Bhatla, N. and Jyoti, K., "A novel approach for heart disease diagnosis using data mining and fuzzy logic", ***International Journal of Computer Applications***, Vol. 54, No. 17, (2012).
19. Dennis, B. and Muthukrishnan, S., "AGFS: Adaptive genetic fuzzy system for medical data classification", ***Applied Soft Computing***, Vol. 25, (2014), 242-252.
20. Darvishi, A. and Hassanpour, H., "A geometric view of similarity measures in data mining", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 28, No. 12, (2015), 1728.
21. Esmaileyan, Z. and Marvi, H., "A database for automatic persian speech emotion recognition: Collection, processing and evaluation", ***International Journal of Engineering***, Vol. 27, (2013), 79-90.
22. Fei, S.-w., "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine", ***Expert Systems with Applications***, Vol. 37, No. 10, (2010), 6748-6752.
23. Qazi, M., Fung, G., Krishnan, S., Bi, J., Rao, R. B. and Katz, A. S., "Automated heart abnormality detection using sparse linear classifiers", ***IEEE Engineering in Medicine and Biology Magazine***, Vol. 26, No. 2, (2007), 56-63.
24. Polat, K., Sahan, S. and Gunes, S., "Automatic detection of heart disease using an artificial immune recognition system (airs) with fuzzy resource allocation mechanism and K-NN (nearest neighbour) based weighting preprocessing", ***Expert Systems with Applications***, Vol. 32, No. 2, (2007), 625-631.
25. Conforti, D., Costanzo, D. and Guido, R., "Medical decision making: A case study within the cardiology domain", ***Journal on Information Technology in Healthcare***, Vol. 5, No. 6, (2007), 343-356.
26. Anbarasi, M., Anupriya, E. and Iyengar, N., "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", ***International Journal of Engineering Science and Technology***, Vol. 2, No. 10, (2010), 5370-5376.
27. Vijaya, K., Khanna Nehemiah, H., Kannan, A. and Bhuvaneswari, N., "Fuzzy neuro genetic approach for predicting the risk of cardiovascular diseases", ***International Journal of Data Mining, Modelling and Management***, Vol. 2, No. 4, (2010), 388-402.
28. Rajkumar, A. and Reena, G. S., "Diagnosis of heart disease using datamining algorithm", ***Global Journal of Computer Science and Technology***, Vol. 10, No. 10, (2010), 38-43.
29. Babaoglu, I., Baykan, O. K., Aygul, N., Ozdemir, K. and Bayrak, M., "Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization", ***Expert Systems with Applications***, Vol. 36, No. 2, (2009), 2562-2566.
30. Khemphila, A. and Boonjing, V., "Heart disease classification using neural network and feature selection", in Systems Engineering (ICSEng), 21st International Conference on, IEEE, (2011), 406-409.
31. Lahsasna, A., Ainon, R. N., Zainuddin, R. and Bulgiba, A., "Design of a fuzzy-based decision support system for coronary heart disease diagnosis", ***Journal of Medical Systems***, Vol. 36, No. 5, (2012), 3293-3306.
32. Peter, T. J. and Somasundaram, K., "Study and development of novel feature selection framework for heart disease prediction",

International Journal of Scientific and Research Publications, Vol. 2, No. 10, (2012), 1-7.

33. Peter, T. J. and Somasundaram, K., "An empirical study on prediction of heart disease using classification data mining techniques", in Advances in Engineering, Science and Management (ICAESM), International Conference on, IEEE, (2012), 514-518.
34. Muthukaruppan, S. and Er, M., "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease", ***Expert Systems with Applications***, Vol. 39, No. 14, (2012), 11657-11665.
35. Santhanam, T. and Ephzibah, E., "Heart disease classification using PCA and feed forward neural networks", in Mining intelligence and knowledge exploration, Springer, (2013), , 90-99.
36. Patel, S. B., Yadav, P. K. and Shukla, D. D., "Predict the diagnosis of heart disease patients using classification mining techniques", ***IOSR Journal of Agriculture and Veterinary Science***, Vol. 4, No. 2, (2013), 61-64.
37. Subanya, B. and Rajalaxmi, R., "Feature selection using artificial bee colony for cardiovascular disease classification", in Electronics and Communication Systems (ICECS), International Conference on, IEEE, (2014), 1-6.
38. Krishnaraj, N. and Vinothkumar, M. R., "Heart disease prediction using ga and MLBPN", ***Heart Disease***, Vol. 2, No. 4, (2014).
39. Wisaeng, K., "Predict the diagnosis of heart disease using feature selection and k-nearest neighbor algorithm", ***Applied Mathematical Sciences***, Vol. 8, No. 83, (2014), 4103-4113.
40. Nguyen, T., Khosravi, A., Creighton, D. and Nahavandi, S., "Classification of healthcare data using genetic fuzzy logic system and wavelets", ***Expert Systems with Applications***, Vol. 42, No. 4, (2015), 2184-2197.
41. Tan, P.-N., "Introduction to data mining", Pearson Education India, (2006).
42. Shaeiri, Z. and Ghaderi, R., "Modification of the fast global K-means using a fuzzy relation with ^{application} in microarray data analysis", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 25, No. 4, (2012), 283-292.
43. Prati, R. C., Batista, G. E. and Monard, M. C., "A survey on graphical methods for classification predictive performance evaluation", ***IEEE Transactions on Knowledge and Data Engineering***, Vol. 23, No. 11, (2011), 1601-1618.
44. Riano, D., Bohada, J. A., Collado, A. and Lopez-Vallverdu, J. A., "MPM: A knowledge-based functional model of medical practice", ***Journal of Biomedical Informatics***, Vol. 46, No. 3, (2013), 379-387.
45. Yan, H., Zheng, J., Jiang, Y., Peng, C. and Li, Q., "Development of a decision support system for heart disease diagnosis using multilayer perceptron", in Circuits and Systems, ISCAS'03. Proceedings of the International Symposium on, IEEE. Vol. 5, (2003), 709-712.
46. Das, R., Turkoglu, I. and Sengur, A., "Effective diagnosis of heart disease through neural networks ensembles", ***Expert Systems with Applications***, Vol. 36, No. 4, (2009), 7675-7680.
47. Chu, C.-M., Chien, W.-C., Lai, C.-H., Bludau, H. B., Tschai, H.-J., Pai, L., Hsieh, S.-M., Chu, N. F., Klar, A. and Haux, R., "A bayesian expert system for clinical detecting coronary artery disease", ***Journal of Medical Sciences***, Vol. 29, No. 4, (2009), 187-194.
48. Pal, D., Mandana, K., Pal, S., Sarkar, D. and Chakraborty, C., "Fuzzy expert system approach for coronary artery disease screening using clinical parameters", ***Knowledge-Based Systems***, Vol. 36, (2012), 162-174.
49. Sanz, J. A., Galar, M., Jurio, A., Brugos, A., Pagola, M. and Bustince, H., "Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system", ***Applied Soft Computing***, Vol. 20, (2014), 103-111.
50. Dash, T., Nayak, S. K. and Behera, H., Hybrid gravitational search and particle swarm based fuzzy mlp for medical data classification, in Computational intelligence in data mining, Vol. 1, (2015), Springer, 35-43.
51. Kumar, R. G., "Performance analysis of soft computing techniques for classifying cardiac arrhythmia", ***Indian Journal of Computer Science and Engineering (IJCSE), ISSN***, Vol. 4, No.6, 459-465.
52. Safdarian, N., Dabanloo, N. J. and Attarodi, G., "A new pattern recognition method for detection and localization of myocardial infarction using t-wave integral and total integral as extracted features from one cycle of ecg signal", ***Journal of Biomedical Science and Engineering***, Vol. 7, No. 10, (2014), 818-824.
53. Abuhasel, K. A., Iliyasu, A. M. and Fatichah, C., A combined adaboost and newfm technique for medical data classification, in Information science and applications, (2015), Springer, 801-809.
54. Nguyen, T., Khosravi, A., Creighton, D. and Nahavandi, S., "Medical data classification using interval type-2 fuzzy logic system and wavelets", ***Applied Soft Computing***, Vol. 30, (2015), 812-822.
55. Pandey, A. K., Pandey, P. and Jaiswal, K., "A heart disease prediction model using decision tree", ***IUP Journal of Computer Sciences***, Vol. 7, No. 3, (2013).
56. Masethe, H. D. and Masethe, M. A., "Prediction of heart disease using classification algorithms", in Proceedings of the world congress on engineering and computer science. Vol. 2, (2014), 22-24.
57. Kim, J., Washio, T., Yamagishi, M., Yasumura, Y., Nakatani, S., Hashimura, K., Hanatani, A., Komamura, K., Miyatake, K. and Kitamura, S., "A novel data mining approach to the identification of effective drugs or combinations for targeted endpoints—application to chronic heart failure as a new form of evidence-based medicine", ***Cardiovascular drugs and therapy***, Vol. 18, No. 6, (2004), 483-489.
58. Shouman, M., Turner, T. and Stocker, R., "Using data mining techniques in heart disease diagnosis and treatment", in Electronics, Communications and Computers (JEC-ECC), Japan-Egypt Conference on, IEEE, (2012), 173-177.

Analysis of Pre-processing and Post-processing Methods and Using Data Mining to Diagnose Heart Diseases

H. Hamidi, A. Daraei

Department of Industrial Engineering, Information Technology Group, K. N.Toosi University of Technology, Tehran, Iran

P A P E R I N F O

Paper history:
Received 27 March 2016
Received in revised form 30 April 2016
Accepted 02 June 2016

Keywords:
Data Mining
Heart Disease
Diagnosis
Prognosis
Treatment

چکیده

امروز، در زمینه پزشکی حجم زیادی داده تولید می شود. کسب دانش مفید از این داده های خام، نیازمند پردازش داده ها و تشخیص الگوهای بامعنی می باشد و این هدف با استفاده از داده کاوی قابل دستیابی است. استفاده از داده کاوی برای تشخیص و پیش بینی بیماری های قلبی، در سال های اخیر یکی از زمینه های مورد علاقه محققان بوده است. در این مطالعه، مروری بر کاربرد الگوریتم های دسته بندی در بیماری های قلبی انجام شده است. مطالعه حاضر، تلاشی برای ارزیابی مطالعات انجام شده در این زمینه می باشد، به طوریکه نتایج حاصل از این بررسی می تواند به دید واضح و روشنی برای مطالعات آینده منجر شود. در ابتدا، وظایف اصلی پزشکی مشخص شده و پس از آن، هر مقاله بر اساس این وظایف بررسی شده است. در نهایت، نتایج از نظر فراوانی استفاده از الگوریتم های دسته بندی، روش های پیش پردازش و پس پردازش، ارائه شده اند. در این مطالعه، ٤٩ مقاله از مطالعات مشابه و با موضوعات مرتبط (از سال۲۰۰۳ تا ۲۰۱۵) جمع آوری و بررسی شده است. بدیهی است که تعداد مقالات در مورد استفاده از الگوریتم های دسته بندی در بیماری های قلبی بسیار قابل توجه است، بنابراین، بررسی همه آنها در مطالعه حاضر، غیر ممکن می باشد. امید است که نتایج این مطالعه، مسیری برای پیشرفت های تحقیقاتی بیشتر در آینده در این حوزه فراهم نماید.

***doi**: 10.5829/idosi.ije.2016.29.07a.06*