International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 13337-13345
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

A Survey to Identify an Efficient Classification Algorithm for Heart Disease Prediction

G. Manikandan*, Aravind Vasudev, Anitha Balasubramanian
School of Computing, SASTRA Deemed University, Thanjavur, India
*Corresponding Author

ABSTRACT:

Classification is one of the prominent data mining techniques. The objective of the classification algorithms is to place the data in the appropriate class. Data mining plays a vital role in medical diagnosis. The aim of this paper is to identify an efficient classification algorithm for cardiovascular disease prediction. The efficiency of each classification algorithm is expressed using two parameters namely accuracy and Root Mean Square Error (RMSE). From our experimental analysis, we infer that iterative classifier optimizer algorithm results in higher accuracy.

KEYWORDS: classification, data mining, cardiovascular disease, iterative classifier optimizer, accuracy, root mean square error.

INTRODUCTION:

The role of data mining in the medical industry has become inevitable [1-10]. Doctors rely on various computer models that make use ofinbuilt data mining algorithms to predict various diseases in patients [11-20]. So, the need of the hour is to identify an efficient algorithm for disease prediction. Therefore, it is essential to design and develop the algorithm with minimal errors [21-34].

In this paper, we have chosen six categories of classification techniques namely bayes, functions, lazy, meta, rules and trees as shown in Figure 1. The algorithms that are used under these categories are Naïve Bayes, logistics, IBK, Iterative Classifier Optimizer, Decision Table, Random Forest. These algorithms are individually applied on Cleveland Heart Disease dataset. The accuracy and RMSE for the classification algorithms after applying on the dataset are determined and analyzed.

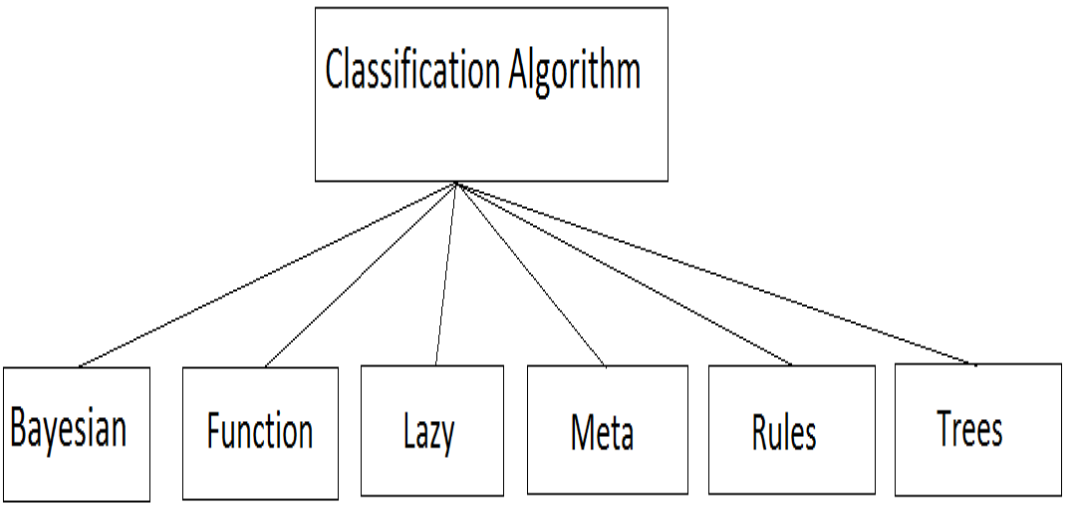

Figure 1- Categories of theclassification algorithm

MATERIALS AND METHODS:

Cardiovascular diseases are one of the main life-threatening diseases. Approximately 12 million deaths occur worldwide due to cardiovascular problems. A raised blood pressure is seen among one in three adults according to world health organization. Heart disease has led to major deaths in India. The World Health Statistics 2012 enlightens that India occupies the 39th position of all the countries suffering from cardiovascular disease. The population suffering heart disease was more in rural areas than in urban areas till 2010, but it has been vice versa since 2015. Thus, there is a serious variance in the population suffering from coronary heart disease in rural and urban areas. In India, population under the age group between 40-49 suffer intensively from heart disease. It is to be noted that the population suffering from cardiovascular disease has doubled in just one decade. Hence heart disease is one of the main reasons that results in fatality in both men and women.

NAÏVE BAYES CLASSIFIER:

This algorithm is a supervised learning method, which is used for classification and can resolve diagnostic and predictive problems based on some historical data. Many learning algorithms can be easily understood and evaluated using Naïve Bayes. It calculates clear probabilities for supposition and remains unaffected by noise in input data. It uses prior, likelihood and posterior as the three main concepts to predict an occurrence of an event, where prior: past experience, likelihood: chance of an event to occur, posterior: prediction of occurrence of an event. The output of this algorithm is computed using the formula given in equation 1.1

posterior = (prior * likelihood) / evidence **1.1**

LOGISTIC:

This class uses multiple classification logistic regression model along with a ridge estimator for building and learning. This regression model uses the logistic function, also called softmax function whose implementation is opaque, to measure the connection between one or more independent variables and categorical dependent variables that are used to predict the target class of the object.The formula for logistic regression and softmax function is given in the equation 1.2 and 1.3. The predicted value obtained from both the equations will always be 0 or 1.

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x) **1.2**

Where, y is the predicted result, b0 is the bias term, b1 is the coefficient and x is the input value.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \; for\; j = 1, \ldots, k$$

1.3

IBK:

This algorithm uses a non-parametric method for regression as well as classification. It performs the k-nearestneighbor algorithm. The majority vote of its neighbors is taken into consideration for classifying an object. The class that is common among its k-nearest neighbors is chosen and the object is assigned to that class. If k is positive and a small integer(say k=1), then the class of that single nearest neighbor is chosen and the object is simply assigned.

K-NN is the simplest among all machine learning algorithms. It locally approximates the function and performs classification. The computation stops only after classification is completely performed, and hence this algorithm is also called "lazy learning method". Here,weights are assigned to the neighbor's contributions so they contribute more than the distant ones.

ITERATIVE CLASSIFIER OPTIMIZER:

Iterative Classifier Optimizer neural network compares the known actual classification of the record with their classification of the record. The algorithm is modified for further iterations by feeding back the errors obtained from the classification of the first record. This algorithm works like a crude electronic network of neurons and hence can be compared to the brain's neural structure.

In iterative classifier optimizer, the records are dispensed one at a time to the network. The process is often repeated after all the input cases are presented and hence this algorithm is a key property of artificial neural network. It can be trained for a particular application by structuring the neural network. The training begins by choosing the initial weights randomly.

DECISION TABLE:

Decision Table is derived from the decision tree. The action performed is based on theconditions that are provided for the decision making. Decision table can be used when there is a consistent number of a condition to be checked. The action must be performed on every single node of the tree, even if it is analyzed to be true. A decision tree can hold more conditions in one branch to be assessed. The templates and data are used for the decision tree; it also consists of rows and columns. Similarly, rules within a business are expressed in templates; each row independently collects the data and stores it separately. The combination of data results in a template to generate a new rule, if the set of templates does not follow the rules then, Decision tables cannot be generated.

RANDOM FOREST:

From the training subset of data, Random Forest produces a set of decision trees. This algorithm determines the class of an object by combining the votes from desperate decision trees. Since the algorithm uses votes from different decision trees, this classification technique works well and is less vulnerable to noise. Let the training subset of data be [A1, A2, A3, A4] with labels [C1, C2, C3, C4]. Random forest takes the training subset as input and produces three decision trees. For example, 1. [A1, A2, A3] 2. [A1, A2, A4] 3. [A2, A3, A4]. The prediction is done depending on the majority of votes obtained from the individual decision tree.

RESULTS AND DISCUSSION:

Dataset Description

Cleveland Heart Disease Dataset obtained from UCI machine learning repository is used as input for the above algorithms. The dataset contains 76 attributes, the dataset is then pre-processed which results in a dataset containing 14 attributes and 303 instances. The dataset even after pre-processing contains a few unknowns which are removed manually. The number of instances after removing unknowns are 297. It is then applied to Weka tool experimenter. The accuracy and RMSE values are then determined and analyzed.

Attribute Description

The attributes that are used in the above algorithms for heart disease prediction are: CP (chest pain type), fbs (fasting blood glucose), age, trestbps (resting blood pressure), exang (exercise-induced angina), thal (genetic disorders), sex, old peak, restecg (rest electrocardiogram), CA (major vessels coloured by fluoroscopy), slope, thalach (max heart beat rate achieved), cholesterol, num (status). Diagnosis of heart disease is a complicated task that needs skill and vast knowledge in that domain. It depends on the doctor's experience and the current health condition of the patient in many cases.

Weka

It is a software issued under the General Public License GNU which can be easily downloaded and used from the internet. It contains a variety of machine learning algorithms using which several data mining tasks can be performed. It is a collection of tools for clustering, classification, association, rules, pre-processing, visualization and regression. It is also widely used for developing new machine learning algorithms. Table I summarizes the accuracy and RMSE values for the above six algorithms.

Table 1 - Accuracy and RMSE values

Algorithm	Accuracy	RMSE
bayes.NaiveBayes	83.5017	0.3643
functions.Logistic	83.5017	0.3515
lazy.IBK	76.431	0.4837
meta.IterativeClassifierOptimizer	84.1751	0.363
rules.DecisionTable	81.4815	0.377
trees.RandomForest	83.8384	0.355

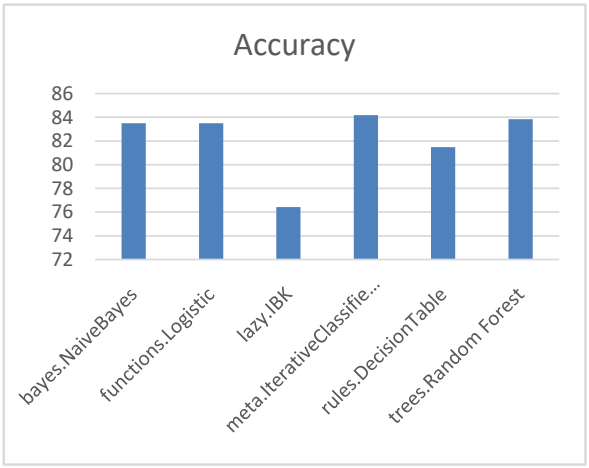

Figure 2 - Accuracy values for Different Classifiers

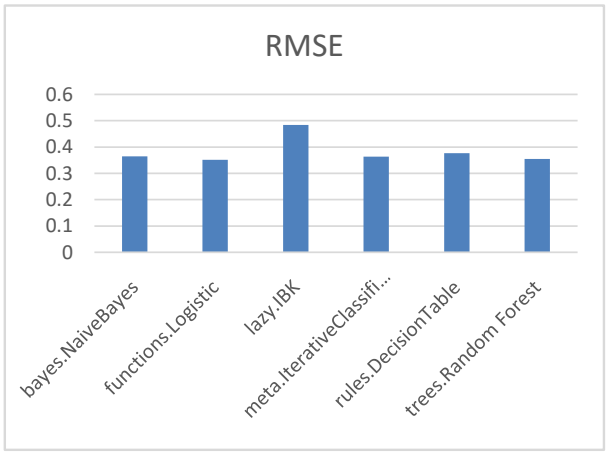

Figure 3 - RMSE values for Different Classifiers

Figure 2 and Figure 3 shows the graphical representation of accuracy and RMSE values obtained when the values in the dataset are applied to the above-mentioned algorithms.

CONCLUSION:

The objective of this paper is to identify an efficient classification algorithm for heart disease prediction. From the graph (Figure 2 and Figure 3), Iterative Classifier Optimizer algorithm under the meta class gives highest accuracy and Decision Table algorithm under rules class has the highest RMSE value. We can also infer that Iterative Classifier Optimizer algorithm produces the highest accuracy of all the classification algorithms available in the Weka tool experimenter for the chosen dataset. In future, the ensemble of different combinations of algorithms under different classes of Weka tool can be constructed to get better accuracy.

REFERENCES:

[1] A. K. S, Dr. D. P. Shukla. A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level. International Journal of Engineering and Computer Science. 2013 Sep 30;2(9):1663-71
[2] Randa EI Bialy, Mostafa A. Salama, Omar Karam. An ensemble model for heart disease data sets: a generalized model. International Conference on Informatics and Systems. 2016 May 11;2(76):191-6
[3] R. Setthukkarase, Kannan. An Intelligent System for mining Temporal rules in Clinical database using Fuzzy neural network. European Journal of Scientific Research. 2012 Feb 15;3(26):1167-78
[4] M.A.Jabbar, Shirina Samreen. Heart Disease prediction system based on hidden naïve bayes classifier. International Conference on Circuits, Controls, Communications and Computing(I4C). 2016 Oct 02;3(62):1-5
[5] Shadab Adam Pattekari and Asma Parveen. Prediction System for Heart Disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences. 2012 Jun 12;3(3):290-4
[6] R.Suganya, S.Rajaram, A.Sheik Abdullah, V.Rajendran. A novel feature selection method for prediction of heart diseases with data mining techniques. Asian Journal of Information Technology. 2016 Jan 18;8(15):1314-21
[7] Nilakshi P. Waghulde, Nilima P. Patil. Genetic Neural Approach for Heart Disease Prediction. International Journal of Advanced Computer Research. 2014 Sep 16;4(3):778-784
[8] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni. Predictive data mining for medical diagnosis: An overview of Heart disease prediction. International Journal of Computer Applications. 2011 Mar 08;17(8):43-8
[9] M. Anbarasi, E. Anupriya, N.CH.S.N. Iyengar. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology. 2010 Oct 14;2(10):5370-5376
[10] Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Gao, X. Z., & Indragandhi, V. (2017). A hybrid quantum-induced swarm intelligence clustering for the urban trip recommendation in smart city. Future Generation Computer Systems, 83, 653-673.
[11] Subramaniyaswamy, V., & Logesh, R. (2017). Adaptive KNN based Recommender System through Mining of User Preferences. Wireless Personal Communications, 97(2), 2229-2247.
[12] Logesh, R., & Subramaniyaswamy, V. (2017). A Reliable Point of Interest Recommendation based on Trust Relevancy between Users. Wireless Personal Communications, 97(2), 2751-2780.
[13] Logesh, R., & Subramaniyaswamy, V. (2017). Learning Recency and Inferring Associations in Location Based Social Network for Emotion Induced Point-of-Interest Recommendation. Journal of Information Science & Engineering, 33(6), 1629–1647.
[14] Subramaniyaswamy, V., Logesh, R., Abejith, M., Umasankar, S., & Umamakeswari, A. (2017). Sentiment Analysis of Tweets for Estimating Criticality and Security of Events. Journal of Organizational and End User Computing (JOEUC), 29(4), 51-71.
[15] Indragandhi, V., Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Siarry, P., & Uden, L. (2018). Multi-objective optimization and energy management in renewable based AC/DC microgrid. Computers & Electrical Engineering.
[16] Subramaniyaswamy, V., Manogaran, G., Logesh, R., Vijayakumar, V., Chilamkurti, N., Malathi, D., & Senthilselvan, N. (2018). An ontology-driven personalized food recommendation in IoT-based healthcare system. The Journal of Supercomputing, 1-33.

[17] Arunkumar, S., Subramaniyaswamy, V., & Logesh, R. (2018). Hybrid Transform based Adaptive Steganography Scheme using Support Vector Machine for Cloud Storage. Cluster Computing.
[18] Indragandhi, V., Subramaniyaswamy, V., & Logesh, R. (2017). Resources, configurations, and soft computing techniques for power management and control of PV/wind hybrid system. Renewable and Sustainable Energy Reviews, 69, 129-143.
[19] Ravi, L., & Vairavasundaram, S. (2016). A collaborative location based travel recommendation system through enhanced rating prediction for the group of users. Computational intelligence and neuroscience, 2016, Article ID: 1291358.
[20] Logesh, R., Subramaniyaswamy, V., Malathi, D., Senthilselvan, N., Sasikumar, A., & Saravanan, P. (2017). Dynamic particle swarm optimization for personalized recommender system based on electroencephalography feedback. Biomedical Research, 28(13), 5646-5650.
[21] Arunkumar, S., Subramaniyaswamy, V., Karthikeyan, B., Saravanan, P., & Logesh, R. (2018). Meta-data based secret image sharing application for different sized biomedical images. Biomedical Research,29.
[22] Vairavasundaram, S., Varadharajan, V., Vairavasundaram, I., & Ravi, L. (2015). Data mining-based tag recommendation system: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(3), 87-112.
[23] Logesh, R., Subramaniyaswamy, V., & Vijayakumar, V. (2018). A personalised travel recommender system utilising social network profile and accurate GPS data. Electronic Government, an International Journal, 14(1), 90-113.
[24] Vijayakumar, V., Subramaniyaswamy, V., Logesh, R., & Sivapathi, A. (2018). Effective Knowledge Based Recommeder System for Tailored Multiple Point of Interest Recommendation. International Journal of Web Portals.
[25] Subramaniyaswamy, V., Logesh, R., & Indragandhi, V. (2018). Intelligent sports commentary recommendation system for individual cricket players. International Journal of Advanced Intelligence Paradigms, 10(1-2), 103-117.
[26] Indragandhi, V., Subramaniyaswamy, V., & Logesh, R. (2017). Topological review and analysis of DC-DC boost converters. Journal of Engineering Science and Technology, 12 (6), 1541–1567.
[27] Saravanan, P., Arunkumar, S., Subramaniyaswamy, V., & Logesh, R. (2017). Enhanced web caching using bloom filter for local area networks. International Journal of Mechanical Engineering and Technology, 8(8), 211-217.
[28] Arunkumar, S., Subramaniyaswamy, V., Devika, R., & Logesh, R. (2017). Generating visually meaningful encrypted image using image splitting technique. International Journal of Mechanical Engineering and Technology, 8(8), 361–368.
[29] Subramaniyaswamy, V., Logesh, R., Chandrashekhar, M., Challa, A., & Vijayakumar, V. (2017). A personalised movie recommendation system based on collaborative filtering. International Journal of High Performance Computing and Networking, 10(1-2), 54-63.
[30] Senthilselvan, N., Udaya Sree, N., Medini, T., Subhakari Mounika, G., Subramaniyaswamy, V., Sivaramakrishnan, N., & Logesh, R. (2017). Keyword-aware recommender system based on user demographic attributes. International Journal of Mechanical Engineering and Technology, 8(8), 1466-1476.
[31] Subramaniyaswamy, V., Logesh, R., Vijayakumar, V., & Indragandhi, V. (2015). Automated Message Filtering System in Online Social Network. Procedia Computer Science, 50, 466-475.

[32] Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Unstructured data analysis on big data using map reduce. Procedia Computer Science, 50, 456-465.
[33] Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Intelligent travel recommendation system by mining attributes from community contributed photos. Procedia Computer Science, 50, 447-455.
[34] Vairavasundaram, S., & Logesh, R. (2017). Applying Semantic Relations for Automatic Topic Ontology Construction. Developments and Trends in Intelligent Technologies and Smart Systems, 48.