Journal of Computer Science

Original Research Paper

Classification of Heart Disease Using Cluster Based DT Learning

[1]Senthilkumar Mohan, [1]Chandrsegar Thirumalai and [2]Abdalah Rababah

[1]School of Information Technology and Engineering, VIT University, Vellore, Tamilnadu, India
[2]Department of Mathematical Sciences, United Arab Emirates University, 15551 Al Ain, UAE

Article history:
Received: 13-04-2019
Revised: 15-07-2019
Accepted: 10-01-2020

Corresponding Author:
Abdalah Rababah
Department of Mathematical Sciences, United Arab Emirates University, 15551 Al Ain, UAE
Email: rababah@uaeu.ac.ae

Abstract: In the rural side, due to the absence of cardiovascular ailment centers, around 12 million people passing away worldwide reported by WHO. The principal purpose of coronary illness is a propensity of smoking. Our Cluster based disease Diagnosis (CDD) applies the ML classifiers to improve the prediction accuracy of cardiovascular diseases. For this we have taken a real Cleveland dataset from UCI. First, the ML performance is evaluated through all features. Then, the dataset is split through the class pairs through its distribution. From this class pair, the significant features are identified through entropy process. Through our CDD approach four significant features are identified from thirteen features. From this four features, the ML performance increases when compared to all other features. That is, in RF model the accuracy improves to 9.5%, SVM by 7.2% and DT model by 2.3%.

Keywords: Classification, Machine Learning

Introduction

In this study, the heart disease processed-dataset has been taken from Cleveland clinic foundation. An investigation says that Coronary illness like Intense Myocardial Dead tissue (AMI), Coronary illness (CHD), Myocardial Localized necrosis (MI) and Cardio Vascular Ailment (CVD) kill one individual at regular intervals in the USA. The expression "cardiovascular ailment" influences the corridors in the heart, which develops with a plaque and ends up restricted, diminishing the stream of blood to the heart. It can prompt chest torment or in the end a heart assault. The extraction of helpful information and mapping of concealed examples and connections from the Cleveland databases, we have to combine distinctive advances. One such is combining information mining with a measurable investigation, machine learning and database innovation. This innovation is utilized as a part of numerous territories including the medicinal administrations. Information mining strategies can be utilized viably in surgical methodology, therapeutic tests, medicine and the revelation of connections among clinical and analysis information alongside anticipating the ailments. The information will be mentioned by the specialists' objective facts and experience. The issue in the choices is that the specialist's ability is not even in each subspecialty and is in a few places as a rare asset.

We apply all 14 features of size 303 samples. The target class includes the distribution from class 0 to 4 where class 0 indicates absence of heart disease and from class 1 to 4 indicates the presence. Based on Machine Learning (ML) models, the primary attributes of heart disease are considered as cp (chest pain type), thal (normal, fixed defect, reversible defect), ca (number of major vessels), thalach (maximum heart rate) and finally, num (heart disease prediction attribute). Based on the correlation method, the attribute pair, slope and old peak takes 0.61 relation. In this study, we are going to apply ML classifiers such as Decision Tree (DT), Random Forest (RF), support vector machine SVM and Linear Model (LM). We are going to apply our cluster based DT to find its accuracy and error. Finally, we are giving our system results with the existing classifiers.

Related Systems

In this section, the proper related systems are considered. This includes ML classifiers such as decision tree, Random Forest, SVM and linear model.

A. Decision Tree

A call tree uses a treelike model of determinations and their possible results, together with happening results, asset costs and utility (Meriem and Abdelaziz, 2019). It is a system to demonstrate algorithmic approach to decide that exclusively contains restrictive administration articulations. Choice trees are typically

utilized in explore, particularly in call investigation, to help to decide a method conceivably to accomplish an objective, however, are a favored instrument in machine learning. The techniques from root to leaf speak to order runs the show (Mokhtar *et al.*, 2016). In call investigation, tree and along these lines the firmly associated impact chart region unit utilized as a distinct and logical choice help device, wherever the standard estimation of driving elective territory unit figured (Dua and Karra, 2017). Choice trees, impact graphs, utility capacities and diverse call examination apparatuses and procedures region unit educated to school kid understudies in resources of business, wellbeing political economy and general wellbeing and zone unit tests of research or administration science methodologies. For a Cleveland training samples of a data *D*, the decision trees are built through high entropy features. The entropy has the following form:

$$Entropy = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij} \quad (1)$$

B. Random Forest

This ensemble model works by building a shell of call trees and acquiescent the classification. Arbitrary call for trees bent for overfitting to their training set (Othman and Azahari, 2016; Liu *et al.*, 2017). For tree learning, RF applies bagging. For a given data *D*, with target responses *Y* which repeats the bagging from $b = 1$ to *B*. The unseen samples x' are made by averaging the predictions $\sum_{b=1}^{B} f_b(x')$ from every individual tree on a sample of *D*:

$$\hat{f} \frac{1}{B} \sum_{b=1}^{B} f_b \left(x'\right) \quad (2)$$

For each tree, the uncertainity is estimated through the standard deviation:

$$\sigma = \sqrt{\frac{\sum_{b-1}^{B} f.(x') - \hat{f}}{B-1}} \quad (3)$$

C. SVM

In system learning, bolster vector machines square measure directed mastering with associated getting to know calculations that wreck down statistics applied for characterization and multivariate exam. Let the dataset, $D = \left\{ y_i x_i \right\}; i = 1,2\ldots,n$ where, $x_i \in R_n$ represents the i^{th} vector and $y_i \in R^n$ represents the target item. The linear SVM finds the optimal hyperplane, $f(x) = w^T x + b$ where, *w* is a dimensional coefficient vector and *b* is an offset. This is set by solving the succeeding optimization problem:

$$\begin{aligned} &Min_{w,b,\zeta} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \zeta_i \\ &s.t..y_i \left(\mathbf{w}^{\mathbf{T}} \mathbf{x}_{\mathbf{i}} b \right) \geq 1 - \zeta_1, \zeta_1 \geq 0, \forall, \in \{1,2,\ldots,m\} \end{aligned} \quad (4)$$

D. Linear Model

It is a statistical method used to characterize at least one indicator attribute. This model essentially depicts the strength of association between a dependent variable *y* and at least one independent variable X_i.

This model is signified as follows:

$$y = \beta_0 + \sum \left(\beta_i X_i + \in_i \right) \quad (5)$$

Clustering Based Disease Diagnosis

Our CDD method applies entropy on class pair distribution to identify the significant features. For this, the Cleveland heart samples are split into possible class pair sample distribution. This class based significant features helps in prediction accuracy rather whole samples feature importance. From the class pair based significant features, the ML models like DT, RF, SVM and LM performance are evaluated. At last, the comparative result has been made with significant and all features to view its performance improvements. It goes through the following steps:

- *Step 1:* Select all the features f_1, f_2,…, f_n, a target class *t* with its samples belong to Cleveland heart disease dataset, *D*
- *Step 2:* Evaluate the ML performance of prediction accuracy through all features and then read its true positive *tp*, false positive *fp*, false negative *fn*, true negative *tn* and error rate
- *Step 3:* Split the dataset based on the class distributions *c*. That is, separate each dataset *D* into D_0 with class 0 samples, D_1 with class 1 samples and so on
- *Step 4:* Create the possible class pairs D_{ij} from the *step 3*

i = 0; j = i+1;
Do: $D_{ij} = (D_i \,||\, D_j)$
i++
Until: i<c;

- *Step 5:* Apply the *DT* on each class pair D_{ij} of *step 4* and evaluate its performance to read its true positive *tp*, false positive *fp*, false negative *fn*, true negative *tn* and error rate

- *Step 6:* Select extract its decisional features with respect to i^{th} and j^{th} class clustered target
- *Step 7:* Extract the interconnecting features among the pair class (i, j). That is, (i^{th} *Class* $\cap j^{th}$ *Class*) $\rightarrow$ *Significant features*
- *Step 8:* Apply the ML classifiers using significant features to evaluate its performance

Results

The experimental results are made on Intel i5 processor with Windows 10 OS and simulated in R-studio using ML packages. Simulation results are made on raw set by setting the proportion as 70% as training, 15% as testing and 15% as validation. The Cleveland heart dataset in Table 1 contains 303 samples with 13 processed attributes and a target class distribution *c* ranges from 0 to 4. Our CDD model predicts a patient from 13 fields to classify whether it belongs to heart risk or not. The class-0 indicates no heart risk, whereas class-1:4 indicates heart risk. Here, class-1 indicates less severity of heart risk and class 4 indicates high severity.

The error matrix of DT, RF, SVM and LM model are presented in the following Table 2 to 5 respectively. Here, the LM model outperforms other models by attaining the accuracy level of 69% whereas, DT, SVM and RF achieves only 66.70%, 64.30% and 59.50%.

From the ML results of class wise prediction, the SVM model performs well in class-0 prediction with no error as in Table 4. Similarly, in class-1, the LM comparatively performs well with 62.50% as in Table 5; in class-2, all ML models are performed with full error as in Table 2 to V; in class-3, the SVM and LM model attains 66.70% error as in Tables 4 and 5; and finally in class-4, the DT model attains no error as in Table 2.

Now, we apply our CDD approach to improve the prediction accuracy through class distribution based significant features via entropy process. Here the Cleveland heart samples have five classes, ranges from 0:4. From the below Table 6, it is clear that the class – 0 has high samples as comparative to others. Hence the pairwise class samples are made with other class such as D_{01}, D_{02}, D_{03} and D_{04}.

Here, class 0 contains 164 samples, class 1 contains 55 samples, class 2 contains 36 samples, class 3 contains 35 samples and class 4 contains 13 samples. The DT performance results are evaluated on these pair-wise samples and the same is depicted from Table 8 to 9. From this result, it is observed that the error rate decreases with the class pair high risk combinations. That is in D_{01} samples, the overall error rate is 20% as depicted in Table 7, whereas, in D_{04} samples, the overall error rate is only 4.40% as depicted in Table 10.

From D_{01} pairwise samples, the entropy features are selected. That is, on class-0, the *cp* feature and on class-1, *cp*, *thal*, *ca* features are identified as the decisional features through entropy. Similarly, for pairwise samples D_{02}, D_{03} *and* D_{04} the respective significant features are depicted in the following Table 11.

From the above class wise D_{ij} features, the interconnected features are extracted as, (i^{th} *Class* $\cap j^{th}$ *Class*) $\rightarrow$ *cp*, *ca*, *thal and old peak.* Hence, from thirteen features only four significant features are extracted. Now, the ML performance is evaluated using the significant features. When compared to all features performance, the significant features performance gets increases and the same is presented in Table 12. That is, in RF model the accuracy improves by 9.5%, SVM by 7.2% and *DT* model by 2.3%.

Table 1: Cleveland raw-heart samples

Samples	Attributes	Target
303	13	0:4

Table 2: Error matrix for DT with overall error 33.30%

Actual	**0**	**1**	**2**	**3**	**4**	**Error**
0	25	0	1	0	0	3.8
1	5	2	0	1	0	75.0
2	3	0	0	0	1	100.0
3	2	1	0	0	0	100.0
4	0	0	0	0	1	0.0

Table 3: Error matrix of RF with overall error 40.50%

Actual	**0**	**1**	**2**	**3**	**4**	**Error**
0	24	1	1	0	0	7.7
1	5	1	1	1	0	87.5
2	1	1	0	2	0	100.0
3	1	0	2	0	0	100.0
4	0	0	1	0	0	100.0

Table 4: Error matrix of SVM with overall error 35.7%

Actual	0	1	2	3	4	Error
0	26	0	0	0	0	0.0
1	7	0	0	1	0	100.0
2	2	0	0	2	0	100.0
3	1	0	1	1	0	66.7
4	0	0	0	1	0	100.0

Table 5: Error matrix of LM with overall error 31%

Actual	0	1	2	3	4	Error
0	25	1	0	0	0	3.8
1	3	3	0	1	1	62.5
2	1	1	0	2	0	100.0
3	0	0	1	1	1	66.7
4	0	0	0	1	0	100.0

Table 6: Cleveland class distribution

Target class	Sample distributions
0	164
1	55
2	36
3	35
4	13

Table 7: D_{01} with overall error rate 20%

Actual	0	1	Error
0	23	3	11.50
1	3	1	75.00

Table 8: D_{02} with overall error rate 11.10 %

Actual	0	2	Error
0	22	2	8.30
2	1	2	33.3

Table 9: D_{03} with overall error rate 10.70%

Actual	0	3	Error
0	22	2	8.30
3	1	3	25.00

Table 10: D_{04} with overall error rate 4.40%

Actual	0	4	Error
0	21	1	4.50
4	0	1	0.00

Table 11: Class wise DT attributes

Class (i, j)	Decisional attributes	
	i^{th} Class	*j^{th} Class*
0, 1	cp	cp, thal, ca
0, 2	old peak, ca	ca, thal
0, 3	thal	thal, old peak, cp
0, 4	old peak	old peak

Table 12: Comparative result of ML via all and significant features

Models	All attributes	Selected attributes cp, thal, ca, old peak	Error rate efficiency
DT	33.3%	31%	2.3%↑
RF	40.5%	31%	9.5%↑
SVM	35.7%	28.5%	7.2%↑
LM	31%	35.8%	4.8%↓

Conclusion

Identification of significant features contributes significantly in decision making. In addition, the significant features play a vital role in resource constrained devices without compromising accuracy. In this paper, we defined a CDD approach to select the significant features through pair wise class distribution in multi-labels and entropy. We demonstrated that our CDD can address both accuracy improvement and feature selection. Our evaluation of CDD on Cleveland samples selects only four features such as *cp*, *ca*, *thal and old peak* from thirteen features. Moreover, though CDD features, the ML performance increases when compared to all features performance. That is, in RF model the accuracy progresses to 9.50%, SVM by 7.20% and DT model by 2.30%.

Author's Contribution

Dr.M.Senthilkumar: Participated in Liteature survey, algorithm,references section.

Prof.Chandrasegar: Participated in Algorithm, implemntation.

Dr.Abdalah Rababah: Participated in revison, results and discussion, Proof reading.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved

References

Ahmadi, E., G.R. Weckman and D.T. Masel, 2018. Decision making model to predict presence of coronary artery disease using neural network and C5.0 decision tree. J. Am. Intell. Humaniz. Comput., 9: 999-1011.

Al Nuaimi, Z.N.A.M. and R. Abdullah, 2017. Neural network training using hybrid particle move artificial bee colony algorithm for pattern classification. J. Inform. Commun. Technol., 16: 314-334.

Amin, M.S., Y.K. Chiam and K.D. Varathan, 2019. Identification of significant features and data mining techniques in predicting heart disease. Telemat. Inform., 36: 82-93.

Anooj, P., 2011. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules and decision tree rules. Open Comput. Sci., 1: 27-40.

Das, R., I. Turkoglu and A. Sengur, 2009. Effective diagnosis of heart disease through neural networks ensembles. Expert Syst. Applic., 36: 7675-7680.

Dey, A., J. Singh and N. Singh, 2016. Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis. Int. J. Comput. Applic., 140: 27-31.

Dua, D. and T.E. Karra, 2017. UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA.

Dwivedi, G., R.K. Srivastava and S.K. Srivastava, 2018. A generalized fuzzy TOPSIS with improved closeness coefficient. Expert Syst. Applic., 96: 185-195.

Hannan, M., M.R. Islam, M.A. Haque, M.S. Hossain and A. Ulhaq *et al*., 2019. Automated face detection, recognition and gender estimation applied to person identification. J. Comput. Sci., 15: 395-415. DOI: 10.3844/jcssp.2019.395.415

Jiang, W., X. Xing, S. Li, X. Zhang and W. Wang, 2019. Synthesis, characterization and machine learning based performance prediction of straw activated carbon. J. Clean. Prod., 212: 1210-1223.

Jiang, W., X. Xing, X. Zhang and M. Mi, 2019. Prediction of combustion activation energy of NaOH/KOH catalyzed straw pyrolytic carbon based on machine learning. Renew. Energy, 130: 1216-1225.

Kasim, M.M., R. Kashim and S.A.M.N. Khan, 2017. A linear programming based model to measure efficiency and effectiveness of undergraduate programs. J. Inform. Commun. Technol., 16: 394-407.

Krishan, G.K., M. Somanadh, C. Thirumalai and M.S. Kumar. 2018. One-dimension force balance system for hypersonic vehicle an experimental and fuzzy prediction approach. Mater. Today Proc., 5: 13547-13555.

Liu, X., X. Wang, Q. Su, M. Zhang and Y. Zhu *et al*., 2017. A hybrid classification system for heart disease diagnosis based on the RFRS method. Comput. Math. Methods Med., 2017: 1-11.

Meriem, A. and M. Abdelaziz, 2019. Combining model-based testing and failure modes and effects analysis for test case prioritization: A software testing approach. J. Comput. Sci. DOI: 10.3844/jcssp.2019

Mokhtar, S.A., I.W.H. Wan and N.M. Norwawi, 2016. Modeling reservoir water release decision using adaptive neuro fuzzy inference system. J. Inform. Commun. Technol., 15: 141-152.

Mujahid, K.R. and C. Thirumalai, 2017. Pearson Correlation Coefficient Analysis (PCCA) on adenoma carcinoma cancer. Proceedings of the International Conference on Trends in Electronics and Informatics, May 17-17, Tirunelveli, India, pp: 492-495. DOI: 10.1109/ICOEI.2017.8300976

Nahar, J., T. Imam, K.S. Tickle and Y.P.P. Chen, 2013a. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Syst. Applic., 40: 96-104.

Nahar, J., T. Imam, K.S. Tickle and Y.P.P. Chen, 2013b. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst. Applic., 40: 1086-1093.

Nazari, S., M. Fallah, H. Kazemipoor and A. Salehipour, 2018. A fuzzy inference-fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases. Expert Syst. Applic., 95: 261-271.

Othman, M. and S.N.F. Azahari, 2016. Deseasonalised forecasting model of rainfall distribution using fuzzy time series. J. Informat. Commun. Technol., 15: 153-169.

Paul, A.K., P.C. Shill, M.R.I. Rabin and M.A.H. Akhand, 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. Proceedings of the 5th International Conference on Informatics, Electronics and Vision, May 13-14, IEEE Xplore Press, Dhaka, Bangladesh, pp: 145-150. DOI: 10.1109/ICIEV.2016.7759984

Ramli, R., F. Jamaluddin E.M.N.E.A. Bakar, M.Y. Alias and N.I. Mahat *et al*., 2013. Assignment of spectrum demands by merits via analytic hierarchy process and integer programming. J. Inform. Commun. Technol., 12: 39-53.

Rao, S.N., P. Shenoy, M.M. Gopalakrishnan and A.B. Kiran, 2018. Applicability of the Cleveland clinic scoring system for the risk prediction of acute kidney injury after cardiac surgery in a South Asian cohort. Indian Heart J., 70: 533-537.

Sabahi, F., 2018. Bimodal Fuzzy Analytic Hierarchy Process (BFAHP) for coronary heart disease risk assessment. J. Biomed. Inform., 83: 204-216.

Samuel, O.W., G.M. Asogbon, A.K. Sangaiah, P. Fang and G. Li, 2017. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Syst. Applic., 68: 163-172.

Shah, S.M.S., S. Batool, I. Khan, M.U. Ashraf and S.H. Abbas *et al*., 2017. Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. Phys. A Stat. Mech. Applied, 482: 796-807.

Shao, Y.E., C.D. Hou and C.C. Chiu, 2014. Hybrid intelligent modeling schemes for heart disease classification. Applied Soft Comput. J., 14: 47-52.

Sharma, K., B. Muktha, A. Rani and C. Thirumalai, 2017. Prediction of benign and malignant tumor. Proceedings of the International Conference on Trends in Electronics and Informatics, May 11-12, IEEE Xprore Press, Tirunelveli, India, pp: 1057-1060. DOI: 10.1109/ICOEI.2017.8300871

Srisawat, C. and J. Payakpate, 2016. Comparison of MCDM methods for intercrop selection in rubber plantations. J. Inform. Commun. Technol., 15: 165-182.

Thirumalai, C. and K.S. Sree, 2017. Analysis of cost estimation function for Facebook web click data. Proceedings of the International conference of Electronics, Communication and Aerospace Technology, Apr. 20-22, IEEE Xplore Press, Coimbatore, India, pp: 172-175. DOI: 10.1109/ICECA.2017.8212788

Thirumalai, C. and R. Manzoor, 2017a. Cost optimization using normal linear regression method for breast cancer type I skin, Proceedings of the International conference of Electronics, Communication and Aerospace Technology, Apr. 20-22, IEEE Xplore Press, Coimbatore, India, pp: 264-268. DOI: 10.1109/ICECA.2017.8212813

Thirumalai, C. and R. Manzoor, 2017b. Investigating the breast cancer tissue utilizing semi-supervised learning and similarity measure. Proceedings of the International conference of Electronics, Communication and Aerospace Technology, Apr. 20-22, IEEE Xplore Press, Coimbatore, India, pp: 269-274. DOI: 10.1109/ICECA.2017.8212814

Thirumalai, C., A. Duba and R. Reddy, 2017a. Decision making system using machine learning and Pearson for heart attack. Proceedings of the International conference of Electronics, Communication and Aerospace Technology, Apr. 20-22, IEEE Xplore Press, Coimbatore, India, pp: 206-210. DOI: 10.1109/ICECA.2017.8212797

Thirumalai, C., M. Vignesh and R. Balaji, 2017b. Data analysis using box and whisker plot for lung cancer. Innov. Power Adv. Comput. Technol., 2017: 1-6.

Thirumalai, C., P.A. Reddy and Y.J. Kishore, 2017c. Evaluating software metrics of gaming applications using code counter tool for C and C++(CCCC). Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, Apr. 20-22, IEEE Xplore Press, Coimbatore, India. DOI: 10.1109/ICECA.2017.8212790

Thirumalai, C., S. Monica and A. Vijayalakshmi, 2017d. Heuristics prediction of Olympic medals using machine learning. Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, Apr. 20-22, IEEE Xplore Press, Coimbatore, India. DOI: 10.1109/ICECA.2017.8212734

Thirumalai, G.V., S.K.V. Krishna and K.J. Senapathi, 2017e. Prediction of diabetes disease using control chart and cost optimization-based decision. Proceedings of the International Conference on Trends in Electronics and Informatics, May 11-12, IEEE Xplore Press, Tirunelveli, India, pp: 996-999. DOI: 10.1109/ICOEI.2017.8300857

Vivekanandan, T. and N.C.S.N. Iyengar, 2017. Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. Comput. Biol. Med., 90: 125-136.

Wiharto, H.K. and H. Herianto, 2017. Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis. Int. J. Electr. Comput. Eng., 7: 1023-1031.