New Generation Computing (2021) 39:569–581
https://doi.org/10.1007/s00354-021-00124-4

Ensemble Methods for Heart Disease Prediction

Talha Karadeniz[1] · **Gül Tokdemir**[1] · **Hadi Hakan Maraş**[1]

Received: 5 October 2020 / Accepted: 11 February 2021 / Published online: 1 March 2021

Abstract
Heart disease prediction is a critical task regarding human health. It is based on deriving an Machine Learning model from medical parameters to predict risk levels. In this work, we propose and test novel ensemble methods for heart disease prediction. Randomness analysis of distance sequences is utilized to derive a classifier, which is served as a base estimator of a bagging scheme. Method is successfully tested on medical Spectf dataset. Additionally, a Graph Lasso and Ledoit–Wolf shrinkage-based classifier is developed for Statlog dataset which is a UCI data. These two algorithms yield comparatively good accuracy results: 88.7 and 88.8 for Spectf and Statlog, respectively. These proposed algorithms provide promising results and novel classification methods that can be utilized in various domains to improve performance of ensemble methods.

Keywords Randomness test · Ensemble methods · Heart disease prediction · Covariance estimator · Mahalanobis distance · Bagging classifier · Weak classifier

Introduction

Ensemble methods are those classifiers where a collection of base estimators are built to find a final result of the classification. Each output from base machines is collected to form a voting scheme [11, 14]. Bagging and boosting are two major techniques that are used to obtain an ensemble techniques [11]. Two well-known examples of ensemble methods are Random Forests (RF) [7] and Gradient Boosting Trees. First one is an estimator fusing Decision Trees on subsets of training dataset to control over-fitting [40]. Second one is a greedy approximation of a tree collection [15]. Weak classifiers are the

✉ Talha Karadeniz
talhakaradeniz@cankaya.edu.tr

Gül Tokdemir
gtokdemir@cankaya.edu.tr

Hadi Hakan Maraş
hhmaras@cankaya.edu.tr

[1] Yukaryurtçu Mahallesi Mimar Sinan Cad. No:4, Etimesgut 06790, Ankara, Turkey

base estimators of ensemble methods; Decision Trees are widely used in this context. A Decision Tree is an ML model that establishes an induction [33] machine through a set of human-interpretable rules.

Statlog and Spectf are two well-known datasets used for heart disease prediction [24]. Statlog is a dataset with 13 features obtained from medical measurements whereas Spectf has 44 features extracted from tomography images. Hence, in this study, two novel classification algorithms are proposed for these different datasets.

We have followed the experimentation setup given in [24] and it is appropriate to introduce their Chaos Firefly Attribute Reduction and Fuzzy Logic (CAFL) method here: CAFL is highly based on attribute reduction where Rough Set [31] and Chaos Firefly optimization [13] is used. Then a type-2 Fuzzy Logic system is utilized to make classification.

Related Works

Heart Disease prediction is a field where ensemble methods have been successfully applied [4, 5]. On the other hand, in [24], a Fuzzy Logic approach [23] is experimented together with a rough set [17] feature reduction.

Artificial Neural Networks (ANN) are extensively used in the literature; apart from Deep Learning architectures [37], classical Neural Network structures are also employed [10, 12, 19]. There are also hybrid methods such as [25, 43]. Deep Learning architectures are utilized to improve diagnosis activities of Chronic Kidney Disease and Lung Cancer, respectively, in the domain of online clinical decision support systems [20, 21]. [39] used ANN together with Principal Component Analysis to select features before Breast Cancer classification. [38] PCA and Linear Discriminant Analysis (LDA) are applied to select features and ANN to classify the resulting Breast Cancer data.

Support Vector Machines (SVM) is a method based on minimizing structural risk, where linearly non-separable data are implicitly mapped to a higher dimensional one to obtain separability [9]. SVMs are applied to various problems [27]. In the context of Heart Disease Prediction, it is rather used as a helper method to select features [3] or a component of ensembles [28, 34]. [42] integrated fractal image analysis with SVM to classify breast cancer.

Naive Bayes (NB) classifiers assume that features are independent [36] and choose the class maximizing the overall probability. [41] proposed a decision support system using Naive Bayes. [26] conducted experiments on Cleveland dataset and [30] developed a web-based application upon Naive Bayes categorization.

We claim that ensemble methods are still valuable in domain of heart disease prediction and outperforms firefly algorithm of [24].

Methodology

Two classification algorithms are proposed for heart disease prediction on image and medical measurement datasets. First base classifier proposed—Reference Vector Classifier (RVC)—is based on formulating randomness of distance

sequences with respect to a subset of vectors of image data. Taken a vector **u**, namely, an observation from the training set, other observations are investigated whether their class label sequences are 'regular' when sorted according to the distances to **u**. Core idea is that the more non-random the corresponding label sequence is, more valuable it is **u** for classification. This introduces a wide range of alternatives through the selection of randomness tests [2]. A dataset or a dataset domain can be captured more effectively by a specific randomness test.

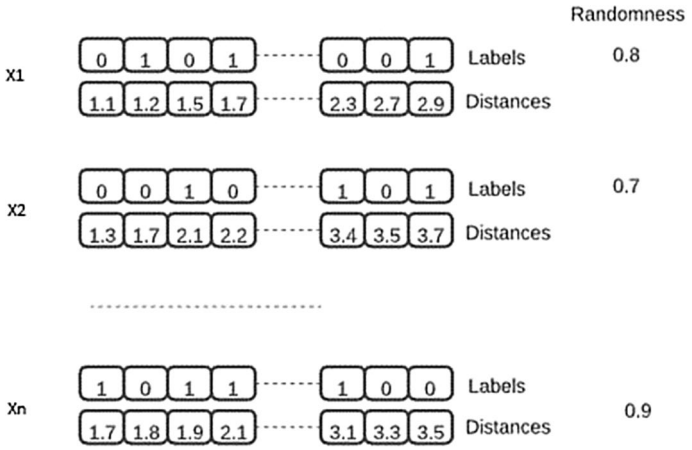

Fig. 1 Class label sequences and distances before sorting

We first present our randomness analysis classifier; for each observation **u**, we find the binary sequence associated with that observation **u**, which is the class

Fig. 2 Class label sequences and distances after sorting

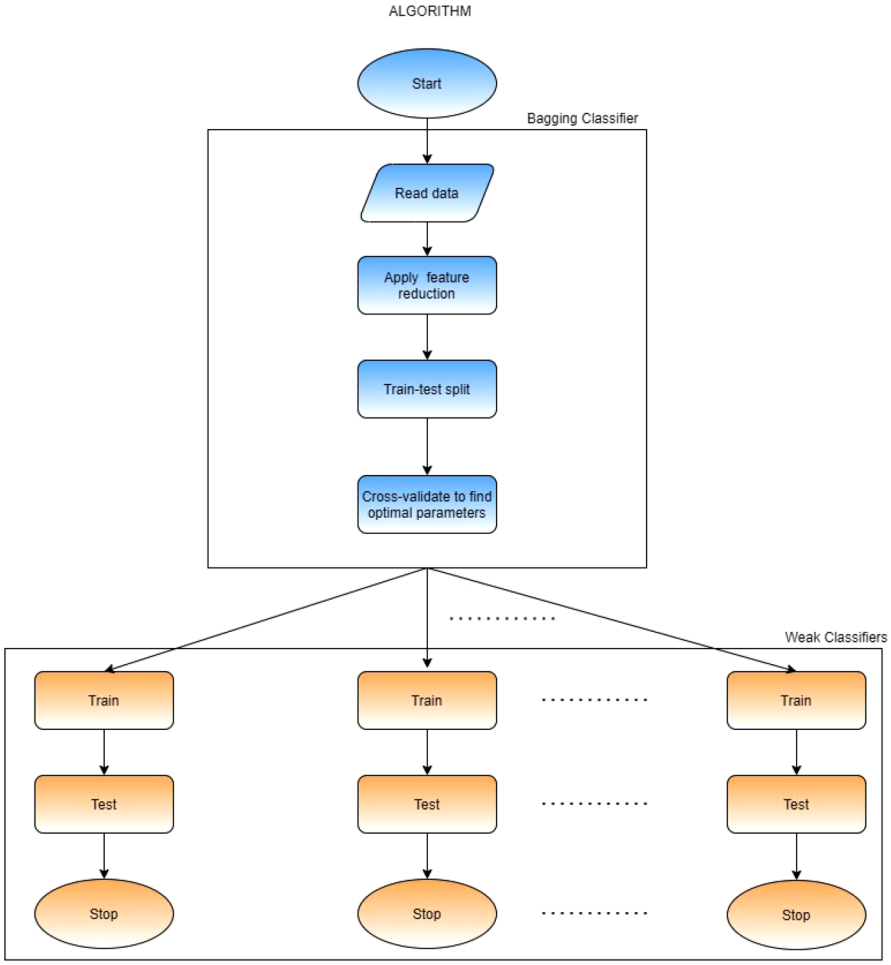

Fig. 3 Overall algorithm

label sequence of other observations when sorted according to their distances to **u**. Randomness calculation of obtained sequence is performed for **u**. More random the sequence is, less important the vector is. This process applied to all vectors in the training set and their label sequences. These label sequences with distances are stored as a matrix which is truncated to form a decision function (Figs. 1, 2, 3).

Algorithm 1 Proposed Method I: Reference Vector Classifier (RVC)

procedure FIT(X, y, n) ▷ Dataset, class labels and number of reference vectors
  $y\_all \leftarrow list()$ ▷ Binary sequence list
  $dist\_all \leftarrow list()$ ▷ Distance measurement list
  $r\_all \leftarrow list()$ ▷ Randomness values for each binary sequence
  $distances \leftarrow euclidean\_distances(X)$ ▷ Matrix for distances between observations
  $i \leftarrow 0$
  while $i < len(X)$ **do**
    $dist \leftarrow distances[i, :]$ ▷ i-th row of distance matrix
    $(dist\_y, dist\_1) \leftarrow extract\_binary\_seq(dist, y)$ ▷ Find binary sequences and distances associated with this observation; these are corresponding labels when distances are sorted
    $r \leftarrow extract\_randomness(dist\_y)$ ▷ Calculate randomness value
    $append(r\_all, r)$
    $append(dist\_all, dist\_1)$ ▷ Append row
    $append(y\_all, dist\_y)$
    $i \leftarrow i + 1$
  end while
  $(ref\_vecs, ref\_labels, ref\_distances) \leftarrow get\_ref\_vecs(X, y\_all, r\_all, dist\_all, n)$ ▷ This is done by sorting according to r values

return $(ref\_vecs, ref\_labels, ref\_distances)$
end procedure

procedure PREDICT($x, n, m, ref\_vecs, ref\_labels, ref\_distances$) ▷ ref_vecs, ref_labels and $ref\_distances$ are results of the FIT() procedure, x is the unseen test observation
  $label\_0 \leftarrow 0$
  $label\_1 \leftarrow 0$
  $i \leftarrow 0$
  while $i < n$ **do**
    $i \leftarrow i + 1$
    $dist \leftarrow euclidean\_distance(x, ref\_vecs_i)$ ▷ Distance of x to i-th reference vector
    $idx \leftarrow binary\_search(dist, ref\_distances_i)$ ▷ Search for closest value in reference distances
    $patch \leftarrow get\_subseq(ref\_labels_i, idx, m)$ ▷ Get subsequence of length $2m + 1$ around idx
    if $sum(patch) < m$ **then**
      $label\_0 \leftarrow label\_0 + 1$
    else
      $label\_1 \leftarrow label\_1 + 1$
    end if
  end while
  $y \leftarrow 0$
  if $label\_0 > label\_1$ **then**
    $ret\_val \leftarrow 0$
  else
    $ret\_val \leftarrow 1$
  end if
return ret_val
end procedure

find_randomness refers to a function for calculating randomness. We have exploited exponential of autocovariance function [29] for randomness calculation. FIT() function returns most important n observations, their label sequences and

distances based on their randomness value. n needs to be determined as a hyperparameter. For observation x, prediction is performed by finding its distance to each reference vector and finding the closest value. The labels of closest values are collected to find the classification result based on majority voting principle. This weak classifier is then plugged into a Bagging Classifier method [6].

Second classifier—Shrunk Covariance Classifier (SCC)—is developed for medical parameter dataset (Statlog) and almost straightforwardly derived from Graphical Lasso [16] and Ledoit–Wolf shrinkage estimation [22], where Glasso and Ledoit–Wolf inverse covariances are fitted and prediction is done with respect to combined Mahalanobis distance. To our knowledge, Glasso and Ledoit–Wolf methods are not applied in this context, that is, in combination with Bagging Classification on heart disease prediction.

Algorithm 2 Proposed Method II: Shrunk Covariance Classifier (SCC)

procedure FIT(X, y) ▷ Training dataset (X) and class labels (y)
  $prec\_0 \leftarrow inv(GraphLasso(X))$ ▷ Fit a Graph Lasso and find inverse
  $prec\_1 \leftarrow inv(Ledoit - Wolf(X))$ ▷ Fit a Ledoit-Wolf and find inverse
  $mean_0 \leftarrow get\_class\_mean(X, y, 0)$ ▷ Retrieve mean for class 0
  $mean_1 \leftarrow get\_class\_mean(X, y, 1)$ ▷ Retrieve mean for class 1
  return $(prec\_0, prec\_1, mean\_0, mean\_1)$
end procedure
procedure PREDICT$(x, prec_0, prec_1, mean_0, mean_1)$ ▷ $prec\_0, prec\_1, mean\_0$ and $mean\_1$ are results of the FIT procedure, x is the test vector
  $a\_0 \leftarrow mahalanobis(x, mean\_0, prec\_0)$ ▷ Find Mahalanobis distance w.r.t. Graph Lasso precision matrix
  $b\_0 \leftarrow mahalanobis(x, mean\_0, prec\_1)$ ▷ Find Mahalanobis distance w.r.t. Ledoit-Wolf precision matrix
  $a\_1 \leftarrow mahalanobis(x, mean\_1, prec\_0)$
  $b\_1 \leftarrow mahalanobis(x, mean\_1, prec\_1)$
  if $a\_0 + 0.25 * b_0 < a\_1 + 0.25 * b_1$ **then**
    $ret\_val \leftarrow 0$
  else
    $ret\_val \leftarrow 1$
  end if

  return ret_val
end procedure

RVC and SCC methods are depicted in detail in Figs. 4, 5 and 6.

Experiments

Experiments are conducted on Spectf and Statlog datasets. First one has 44 features extracted from Single Proton Emission Computed Tomography (SPECT) images. Second one has 13 features: age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative

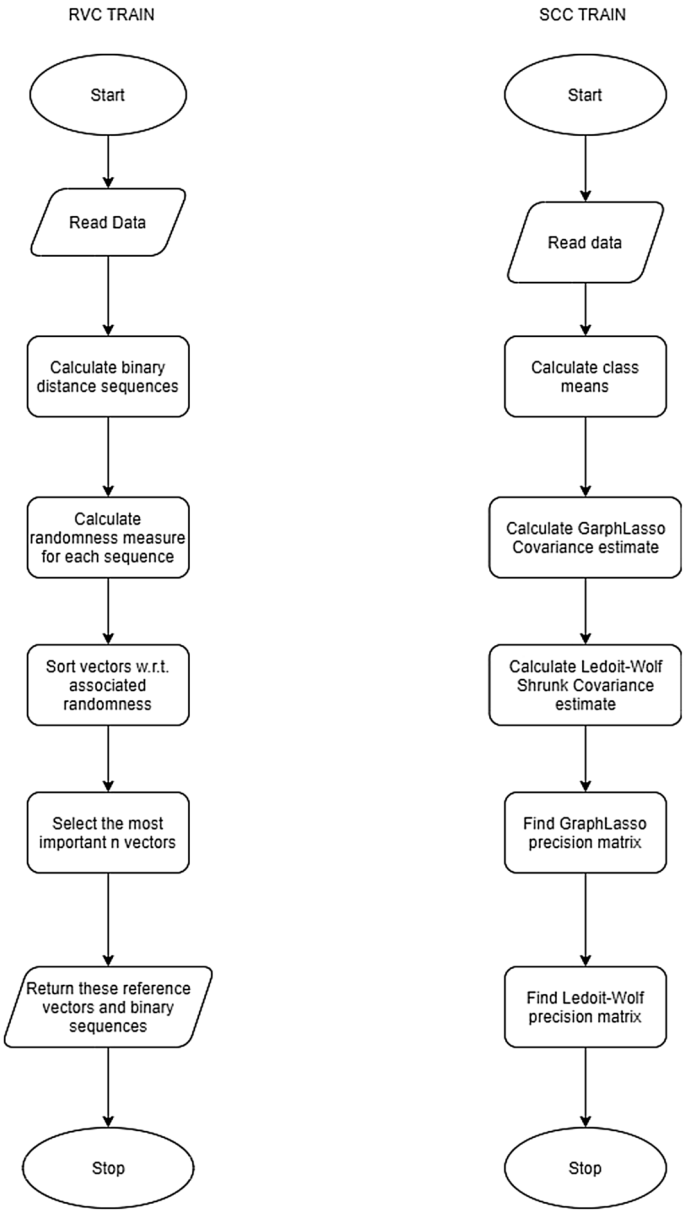

Fig. 4 Base train algorithms

to rest, the slope of the peak exercise ST segment, number of major vessels (0–3) colored by fluoroscopy and defect type.

Experiments are conducted in the same way as [24]; feature reduction is used for both datasets that resulted in 33 features in Spectf and 10 features for Statlog datasets. One third of each dataset is kept for training-validation, remaining for test

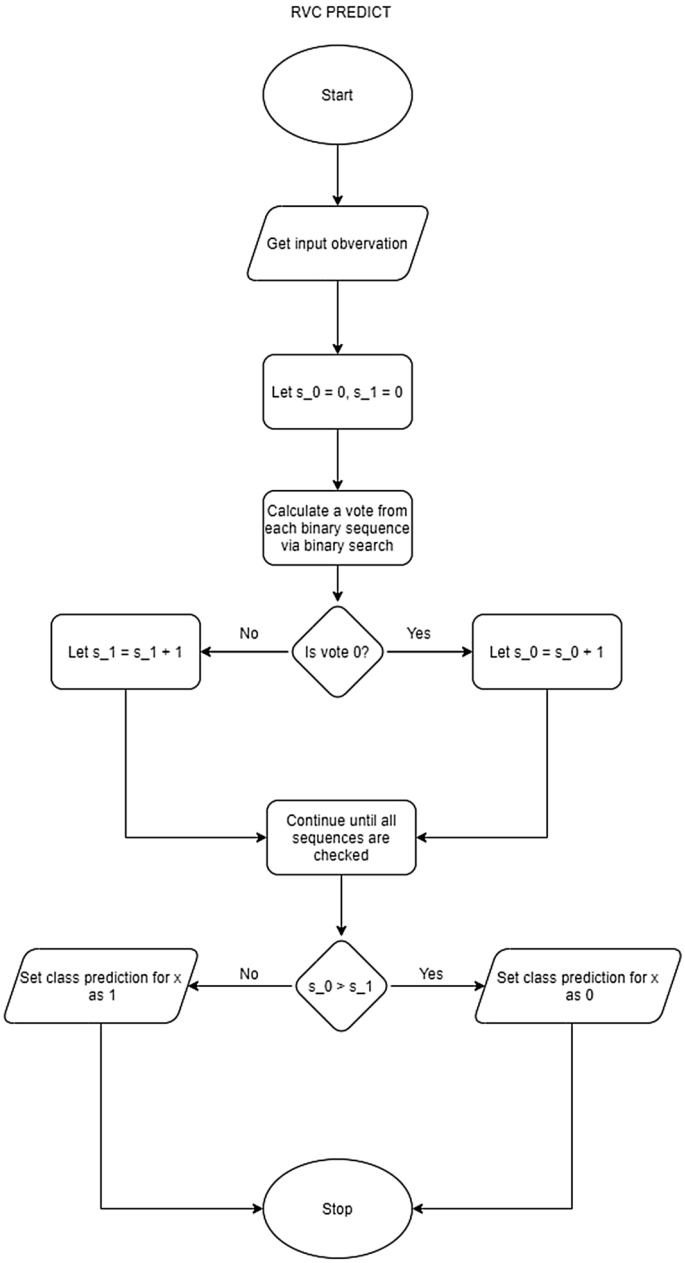

Fig. 5 RVC base predict algorithm

purposes. Cross-validation is applied to decide on the optimal hyperparameters of bagging classifier (i.e. number of estimators, maximum sample size, maximum features etc.). The proposed algorithms' performances are compared with performance

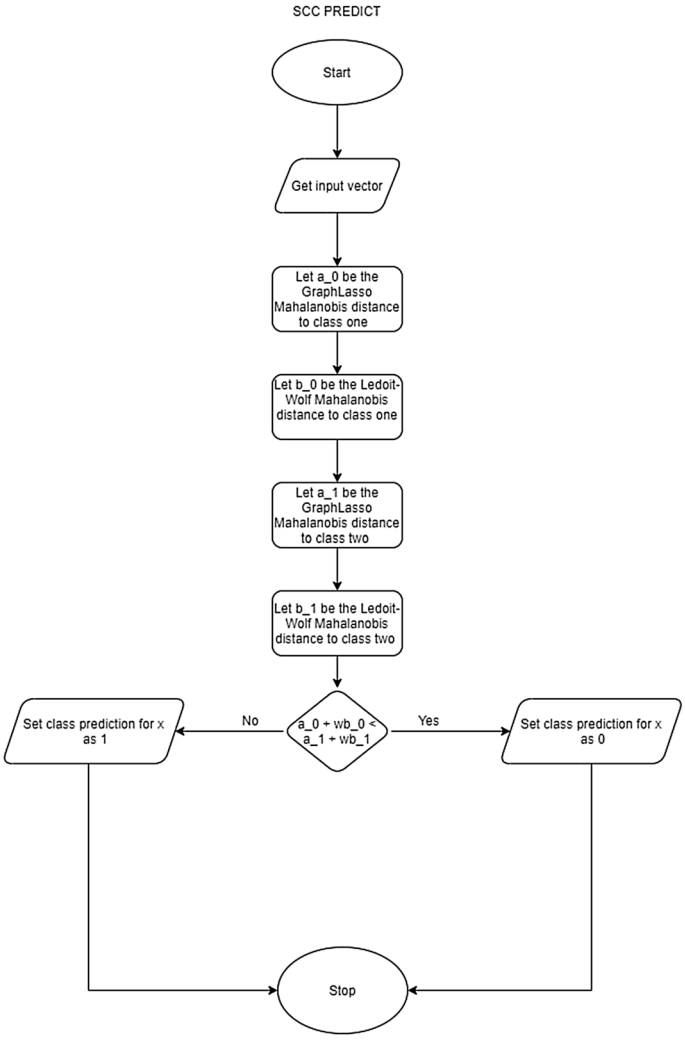

Fig. 6 SCC base predict algorithm

of NB, ANN and SVM, which are obtained from [24]. Scaling to minmax [18] is applied before feature selection and classification. CIFE [8] and ANOVA-based selection are the reduction methods preferred for Spectf and Statlog, respectively.

The algorithm experiments are performed using Python programming language. sklearn [32] is the library used to run Shrunk Covariance estimation, cross-validation and accuracy measurements. Spyder [35] is the IDE where all codes are written.

Additionally, we have conducted a software defect prediction experiment on kc2 dataset where there are 21 features related to software characteristics such as lines of code and McCabe Cyclomatic complexity [1]. In this experiment, we have

compared RVC with RF, SVM and NB. Here, an ANOVA feature reduction (number of reduced features is 10) after Robust Scaling is performed before classification.

Results

To test the performances of the algorithms, several experiments were conducted. We have used four measures: accuracy, precision, recall and f-measure.

Results can be seen from Tables 1 and 2. Our methods outperform classical algorithms and state-of-the-art Chaos Firefly and Fuzzy Logic (CAFL) procedure.

Discussion

The proposed RVC and SCC algorithms outperform CAFL with respect to accuracy metric.

One major advantage of our method over CAFL is that algorithm is still manageable in case of high dimensionality. Other advantage on Statlog is speed; attribute reduction in CAFL takes more than 5 minutes. On the other hand, our total cross-validation, that is whole parameter extraction and testing together with dimension reduction is only about 1 minute. But, one major disadvantage of our methods can be seen when we consider random states. Bagging Classifier implementation has some dependency on random value extraction; we used the optimal solution, giving the maximum test accuracy score. This corresponds somehow 'peeking at the test data' [44]. As a future study, we plan to develop a robust variant of this algorithm.

Second disadvantage in our case is that these classifiers are dataset dependent; that CAFL method itself successful on both Spectf and Statlog.

Our second analysis considers the Shrunk Covariance method, which is a direct application of covariance estimation to classification. This also somehow suffers from curse of dimensionality, but is more straightforward, simpler,

Table 1 Accuracy comparison

Method	Spectf	Statlog
Naive Bayes	79.7	85.2
SVM	79.7	81.5
ANN	77.0	81.5
CAFL	87.2	88.3
RVC, SCC	**88.7**	**88.8**

Table 2 Performance of our methods

Dataset	Precision	Recall	F score
SCC (Statlog)	85.5	89.8	87.6
RVC (Spectf)	91.4	93.7	91.4

interpretable and accurate than CAFL. Of course, Bagging Classifier random state problem arises here, too. Nevertheless, overall speed again is better than CAFL which adds an advantage through the duration of training time.

SVM captures non-linearity via kernel trick. Results indicate that one needs a more sophisticated kernel to derive an accurate classifier on Spectf and Statlog. RVC resolves this by carrying the decision step to distances to specific ('important') observations.

Naive Bayes assumes that features are independent but albeit the shrunk nature of SCC, from results on Statlog, we can see that, in this case, variable interactions can be valuable.

To sum up, proposed algorithms are more accurate and efficient than standard methods and CAFL, the state-of-the art technique in context of heart disease prediction.

Conclusion

In this work, we proposed two algorithms, namely RVC and SCC, for two important datasets, Spectf and Statlog, respectively. We have shown that randomness test-based importance detection is beneficial for classification and shrunk covariance estimators are potentially good as Mahalanobis distance measure sources. Two different feature reduction schemes are plugged into the framework to obtain better accuracy results.

Future Work

Future study will focus on two aspects: first, a more robust variant independent of random states (an average score higher than state of the art) and second, application to various datasets other than heart disease.

Acknowledgements We would like to thank Yusuf "oblomov" Karacaören and Mehmet Fatih "quintall" Karadeniz for their support for conducting the experiments and development of the algorithm.

References

1. Agarwal, S., Tomar, D.: A feature selection based model for software defect prediction. Assessment **65**, (2014)
2. Alcover, P.M., Guillamón, A., Ruiz, M.D.C.: A new randomness test for bit sequences. Informatica **24**(3), 339–356 (2013)
3. Bashir, S., Khan, Z.S., Khan, F.H., Anjum, A., Bashir, K.: Improving heart disease prediction using feature selection approaches. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 619–623. IEEE (2019)

4. Bashir, S., Qamar, U., Khan, F.H., Javed, M.Y.: Mv5: a clinical decision support framework for heart disease prediction using majority vote based classifier ensemble. Arab J Sci Eng **39**(11), 7771–7783 (2014)
5. Bashir, S., Qamar, U., Khan, F.H.: Bagmoov: a novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Austral. Phys. Eng. Sci. Med. **38**(2), 305–323 (2015)
6. Breiman, L.: Bagging predictors. Mach Learn **24**(2), 123–140 (1996)
7. Breiman, L.: Random forests. Mach Learn **45**(1), 5–32 (2001)
8. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res **13**(1), 27–66 (2012)
9. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov **2**(2), 121–167 (1998)
10. Chen, A.H., Huang, S.Y., Hong, P.S., Cheng, C.H., Lin, E.J.: Hdps: Heart disease prediction system. In: 2011 computing in cardiology. IEEE, pp. 557–560 (2011)
11. Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. Springer, Berlin, pp. 1–15 (2000)
12. Durairaj, M., Revathi, V.: Prediction of heart disease using back propagation mlp algorithm. Int J Sci Technol Res **4**(8), 235–239 (2015)
13. Fister, I., Fister, I., Jr., Yang, X.S., Brest, J.: A comprehensive review of firefly algorithms. Swarm Evol Comput **13**, 34–46 (2013)
14. Flennerhag, S.: Introduction to python ensembles (2020). https://www.dataquest.io/blog/introduction-to-ensembles/
15. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232 (2001)
16. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
17. Hu, X., Cercone, N.: Learning in relational databases: a rough set approach. Comput. Intell. **11**(2), 323–338 (1995)
18. Juszczak, P., Tax, D., Duin, R.P.: Feature scaling in support vector data description. In: Proc. asci. Citeseer, pp. 95–102 (2002)
19. Karayılan, T., Kılıç, Ö.: Prediction of heart disease using neural network. In: 2017 International Conference on Computer Science and Engineering (UBMK). IEEE, pp. 719–723 (2017)
20. Lakshmanaprabu, S., Mohanty, S.N., Krishnamoorthy, S., Uthayakumar, J., Shankar, K., et al.: Online clinical decision support system using optimal deep neural networks. Appl. Soft Comput. **81**, 105487 (2019)
21. Lakshmanaprabu, S., Mohanty, S.N., Shankar, K., Arunkumar, N., Ramirez, G.: Optimal deep learning model for classification of lung cancer on ct images. Future Gen. Comput. Syst. **92**, 374–382 (2019)
22. Ledoit, O., Wolf, M.: A well conditioned estimator for large dimensional covariance matrices (2000)
23. Long, N.C., Meesad, P.: An optimal design for type-2 fuzzy logic system using hybrid of chaos firefly algorithm and genetic algorithm and its application to sea level prediction. J. Intell. Fuzzy Syst. **27**(3), 1335–1346 (2014)
24. Long, N.C., Meesad, P., Unger, H.: A highly accurate firefly based algorithm for heart disease prediction. Expert Syst. Appl. **42**(21), 8221–8231 (2015)
25. Malav, A., Kadam, K., Kamat, P.: Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. Int. J. Eng. Technol. **9**(4), 3081–3085 (2017)
26. Medhekar, D.S., Bote, M.P., Deshmukh, S.D.: Heart disease prediction system using naive bayes. Int. J. Enhanced Res. Sci. Technol. Eng. **2**(3) (2013)
27. Moguerza, J.M., Muñoz, A., et al.: Support vector machines with applications. Stat. Sci. **21**(3), 322–336 (2006)
28. Mythili, T., Mukherji, D., Padalia, N., Naidu, A.: A heart disease prediction model using svm-decision trees-logistic regression (sdl). Int. J. Comput. Appl. **68**(16), (2013)
29. Parzen, E.: On spectral analysis with missing observations and amplitude modulation. In: Sankhyā: The Indian Journal of Statistics, Series A, pp. 383–392 (1963)
30. Pattekari, S.A., Parveen, A.: Prediction system for heart disease using naïve bayes. Int. J. Adv. Comput. Math. Sci. **3**(3), 290–294 (2012)

31. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data, vol. 9. Springer, Berlin (2012)
32. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
33. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
34. Radhimeenakshi, S.: Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, pp. 3107–3111 (2016)
35. Raybaut, P.: Spyder-documentation. Available online at: pythonhosted. org (2009)
36. Rish, I., et al.: An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, pp. 41–46 (2001)
37. Rusk, N.: Deep learning. Nat. Methods **13**(1), 35 (2016)
38. Sahu, B., Dash, S., Nandan Mohanty, S., Kumar Rout, S.: Ensemble comparative study for diagnosis of breast cancer datasets. Int. J. Eng. Technol. **7**(4.15), 281–285 (2018)
39. Sahu, B., Mohanty, S., Rout, S.: A hybrid approach for breast cancer classification and diagnosis. EAI Endors. Trans. Scalable Inf. Syst. **6**(20) (2019)
40. sklearn: Random forest classifier (2020). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
41. Subbalakshmi, G., Ramesh, K., Rao, M.C.: Decision support in heart disease prediction system using naive bayes. Indian J. Comput. Sci. Eng. (IJCSE) **2**(2), 170–176 (2011)
42. Swain, M., Kisan, S., Chatterjee, J.M., Supramaniam, M., Mohanty, S.N., Jhanjhi, N., Abdullah, A.: Hybridized machine learning based fractal analysis techniques for breast cancer classification
43. Turabieh, H.: A hybrid ann-gwo algorithm for prediction of heart disease. Am. J. Oper. Res. **6**(2), 136–146 (2016)
44. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.: Experimental comparison of representation methods and distance measures for time series data. Data Min. Knowl. Discov. **26**(2), 275–309 (2013)