Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

An Efficient Svm and Aco-Rf Method for the Cluster-Based Feature Selection and Classification

Kailash Patidar and Dhanraj Verma
Department of Computer Science and Engineering
Dr. A.P.J. Abdul Kalam University, Indore – 452016

Abstract
An efficient support vector machine (SVM) and ant colony optimization-random forest (ACO-RF) method for the cluster-based feature selection and classification were presented and evaluated. It has been evaluated on the breast cancer Wisconsin dataset. The complete process includes preprocessing, feature selection based on class levels clusters and classification. SVM has been applied first for the data classification. It has been applied based on the class labels and further associate the correlations based on the heatmap. Finally, ACO-RF based feature selection mechanism was applied. It will be helpful in the elimination of irrelevant features based on the threshold parameters. The result clearly indicates that SVM and ACO-RF outperforms in all aspects. It is also clear from the heatmap regarding the attribute correlation in all aspects.

Keywords
SVM, ACO, RF, Heatmap

1. Introduction

Data gathering, preprocessing and pattern extraction are the important part of data mining and classification techniques [1-3]. These techniques are useful in the extraction of the meaningful insights [4-6]. So, it can be said that the methods and algorithms of data mining and machine learning algorithms may be useful for computational applicability in different domains [7, 8]. These are used in two ways one is supervised and another in unsupervised way of computing [7−10]. It includes classification and clustering task. Data grouping can be performed through clustering algorithms like k-means, fuzzy c-means (FCM), hierarchical clustering etc. [11, 12]. Another level of pruning and intinction are needed in the complex situation [13]. In the complex problem situations optimization techniques like are ant colony optimization (ACO), particle swarm optimization (PSO), teaching learning-based optimization (TLBO), Cuckoo Search, etc. [14,15] may be used.For the data classification and feature selection machine learning algorithms can be used [16, 17]. Some of the machine learning algorithms are support vector machine (SVM), logistic regression (LR), random forest (RF), naïve Bayes (NB), K-Nearest Neighbor (KNN), etc. [18-21] which can be used for the better classification and clustering.

In 2020, Sebastian et al. [22], tried to characterize the collapse in upper airway for the Obstructive sleep apnoea (OSA) patients. They collected the data of 58 patients who were diagnose with OSA. They recorded the data of audio signal with Microphone (Gooseneck Mini Shotgun). They pre-processed the audio signals and extracted the 50 identical features. They performed the cluster analysis (unsupervised) and using k-means clustering and applied the silhouette analysis for calculating the validation of clusters. They performed the internal and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

external validation of clusters. In this study, they tried to identify the same with the help of unlabeled data. As a limitation, they found the k-means algorithm is sensitive to the initial values and also it is difficult to predict the number of clusters. They achieved an accuracy of 62% and used the silhouette coefficients of 0.79 for the two clusters.In 2020, Alkhafaji et al. [23], They tried to identify the approach of how to clean the data before getting it diagnose for the Heart disease. They considered the data of 300 males and 365 females. They were focused on two steps process. In the initial stage, they performed the cleaning of data with the help of removing the missing data, using the k nearest neighbors'approach and K-means clustering, the noise present in the data and the presence of Inconsistent data. In the second stage, they applied different prediction models like Naïve Bayes Classifier, Decision Tree, and Artificial neural network. they highlighted that undesirable consequences may be occur if clinical decisions are made by experienced physicians instead of applying techniques to hidden data. among all these three prediction models, decision tree outperforms in terms of good performance and accuracy of 98.85%. After that Bayesian classification gave the accuracy of 98.16% and artificial neural network gave the accuracy of 91.31%.In 2020, Shuai et al. [24], analyzed how the national epidemic strategy response to COVID-19 and its severity for the current. They gathered the data from National strategy frame for epidemic and pre-processed it with the help of StandardScaler. After that they performed the cluster analysis by using K-means Clustering. Agglomerative Clustering and DBSCAN Clustering. After performing cluster analysis, they evaluate them on the index of Silhouette Coefficients, Calinski Harabasz Score and Davies Bouldin score. Amongst all they found K-Mean clustering to be the optimal algorithm. They found that some countries are taking precaution measurements and trying to make improvement in immunization of patients.In 2020, Wang et al. [25], they proposed a method of Fiber clustering by adopting the information of structural and functional. They considered the data from Human Connectome Project (HCP Q1) dataset. Initially, they pre-processed the data DTI data and task-fMRI data for obtaining the brain's fibers and the reconstruction of cortical surface took place. Next, this pre-processed data was fed to (CAEEC) Convolutional Auto-Encoders with Embedded Clustering model which generates the clusters of fiber bundles. The main benefit of this joint is to makes the reconstruction easy for features.In 2020, Bu et al. [26], they presented an incremental high-order possibilistic c-means (IHoPCM) algorithm of computing system based on cloud-edge so that they can connect to multiple hospitals and collect the data with the help of co-clustering of medical data. They applied the deep computational model on each hospital so that they can learn the feature tensor on the local edge of computing system. They used these features and upload on cloud so that IHoPCM can be applied to these features and they can make new clustering centers. This is the advantage of their study that they can use cloud to improve the efficiency of clusters. Their study needs to be validated on real time applications.In 2020, Chebanenko et al. [27], they proposed a fuzzy system which can be used to for estimating the patient's level applied for compliance in primary treatment. Their approach was to process and analyses the medical information. This study considers the data of 160 patients who were suffering from cardiovascular disease and they were having the age from 40 years to 85 years. They analyzed the data with the help of fuzzy clustering. They estimated the compliance rate with the help of three criteria i.e. expected efficiency of how they update their way of life, medical therapy and the medical support. They found the three groups of patients. Their results suggest that the outcome received from this can be fed to expert support system, which can use an input for classification. The result of this can be used for individual treatment.In 2020, Doroshenko [28] described the analysis of COVID-19 in Italy. They performed the comparison between two

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

methods i.e. hierarchical and k-Means clustering. They took the sample of 1113 sample entries from Feb, 120 to April 16, 2020. They applied the k-Means clustering first and made the clusters by randomly dividing the data. Next, they applied Hierarchical clustering for building a nest of partitions. They performed a detailed analysis for the clusters obtained. These clusters give information about the region from where patients belong. Further, they suggested to perform more detailed analysis for revealing the dependence of data.In 2020, Kubo et al. [29] proposed a method of designing artificial humeral head model with the help of Kmeans++ and the PCA. They considered the 22 males for this experiment between 18 to 79 years who are not having any disease in their shoulder. Initially, they segmented and did alignment of the Humeral Head. The made alignment after applying the affine transformation. Further, they applied Artificial Humeral Head by using the Kmeans++ and Principal component analysis. They took the k size of 3,4,5 for making the clusters and did evaluation eith the help of leave-one-out cross validation (LOOCV). They found that artificial humeral head model is much similar to the shape of actual humeral head. Further, they decided to consider a greater number of subjects so that they can improve the reliability and also there is a need to establish a criterion for calculating the k values. In 2020. Meniailov et al. [30], discussed the approach of identifying the heart disease of similar characteristics. In this study they performed a task of data clustering. They tried to determine the likelihood of the heart disease for such type of patients who are having same characteristics. They performed the statistical analysis by using the method of Fuzzy C-means and they implemented this with the help of Python and its library i.e. Panda. They analyzed that this statistical analysis is helpful in determining the conditions and also not limited to biomedical field only.In 2020, Roy et al. [31], discussed the methodology to be adopted for identifying the disease in any organ by using Modified-C Clustering technique. They considered the MRI images of eyes, liver and brain. They tried to conclude it higher accuracy in detecting the disease in particular body organ and also with the higher efficiency in detecting the disease. They proposed a classification approach i.e. Modified C-Clustering for detection and the segmentation purpose. They achieved the accuracy more than 98% for the random inputs they applied based on the detection of edge and pattern. As a limitation, they found it to be implement on real time and suggested to use them with artificial intelligence as the future scope.In 2020, Huijuan and Zhenjiang [32], discussed the use of clustering approach in the field of education, medical etc. Particularly they discussed the K-means algo. on SPSS tool for describing the use of data mining platform. They applied the K-Means algorithmic thinking and further applied it with K-Means clustering process. They did this analysis on teaching where 1369 students participated from 16 different branches and 58 learning centers. As a result, they found that as the data is growing rapidly so this approach is helpful in analyzing the data in every field because of its high efficiency and simplicity in use.In 2020, Sudhagar and Renjith [33] discussed the requirements of using clustering techniques for maintain the high-dimensional data available today in every field. They showed the comparison between various data mining techniques and discussed the challenges available with each technique. They proposed approach is helpful in maintaining the high-dimensional records. They used the shift clustering and ensemble clustering approach. They evaluated the proposed approach by using the data of diabetes and breast cancer dataset.

This paper explores the clustering mechanism with the combination of ACO-RF method for improving the classification mechanism.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

2. Materials and Methods

Breast Cancer Wisconsin dataset has been considered from the UCI repository [34]. The following dataset has been used for the experimentation. The number of instances in this dataset are 699. Total number of attributes are 10. Class labels are two.

An efficient SVM and ACO-RF method for the cluster-based feature selection and classification were presented in this paper. It includes feature selection based on the ACO-RF method based on the clusters. The classification has been performed by SVM. The complete approach consists of preprocessing, cluster-based feature selection and classification.

First the data was preprocessed for the missing values and for the unwanted data. Then selection of relevant features was performed. It is important as it may be helpful in accurate classification. It positively affects the performance of the approach. It is also performed to reduce those variables which can slow the development of the model. It is acquired for the inclusion of the most relevant impactful attributes and parameters. The models are then evaluated based on F1-Score, precision, recall, sensitivity and accuracy. The algorithms of these approaches are as follows:

ACO algorithm:
Step 1: Initialization has been performed based on relative importance and trail.
Step 2: Construction of the ant system is started based on the probability state to move into.
Step 3: Finally update the trail based on the movement of ant system.
Step 4: Update the trail matrix.
Step 5: Apply the termination condition in each step.
Step 6: Iterate over 100, 200 and 300 cycles for checking the similar feature-based cluster attribute.
Step 7: If the termination satisfied stop the cycle.
Step 8: Finish.

RF algorithm:

Step1: Initialization has been performed.
Step 2: K data points have been selected randomly. It has been selected from the training set.
Step 3: Gather the selected data points from the associated tree. Based on this build the decision tree.
Step 4: Consider the number for the decision tree construction.
Step 5: Step 2 to 5 will be repeated till the cycle criteria for termination.
Step 6: Check and update the new data points.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

Figure 1 shows the complete mechanism of our approach.

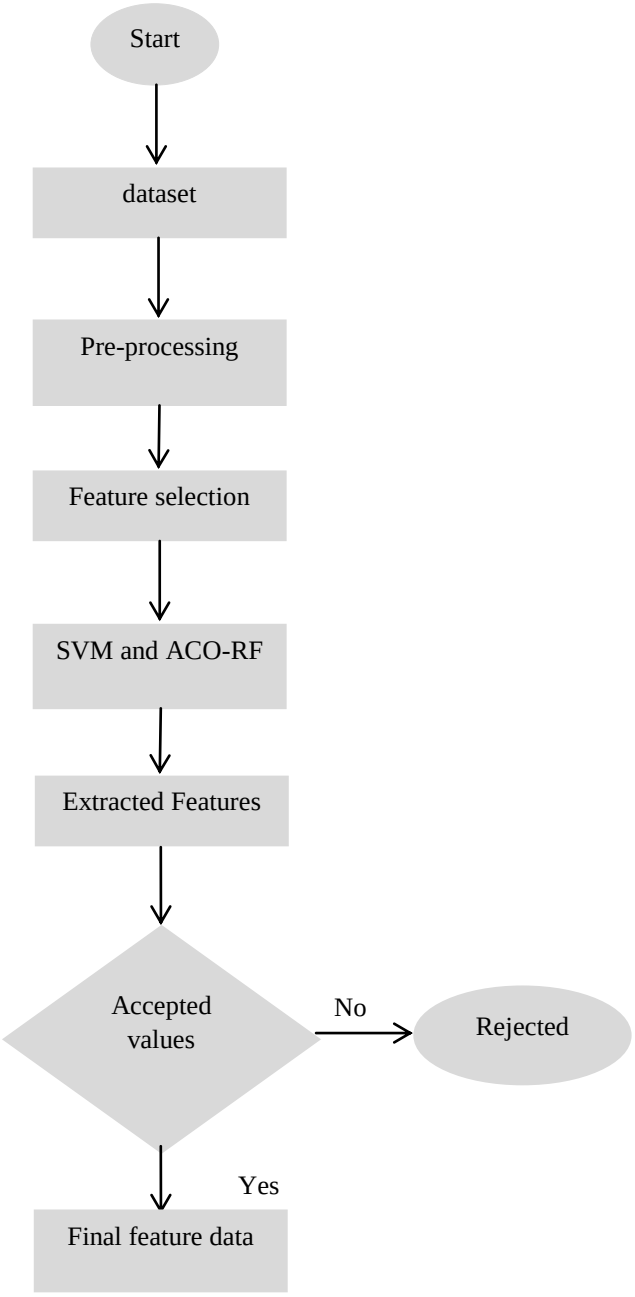

Figure 1: Complete working procedure of the approach

3. Results

This section shows the result based on SVM and ACO-RF approach. The performance evaluation measures are accuracy, precision, recall and F1-Score. Root mean square error (RMSE), mean square error (MSE) and mean absolute error (MAE) were considered for the error rate evaluation. Figure 2 shows theheatmap based data correlation between the attributes based

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

on SVM. Figure 3 shows the heatmap based data correlation between the attributes based on ACO-RF.

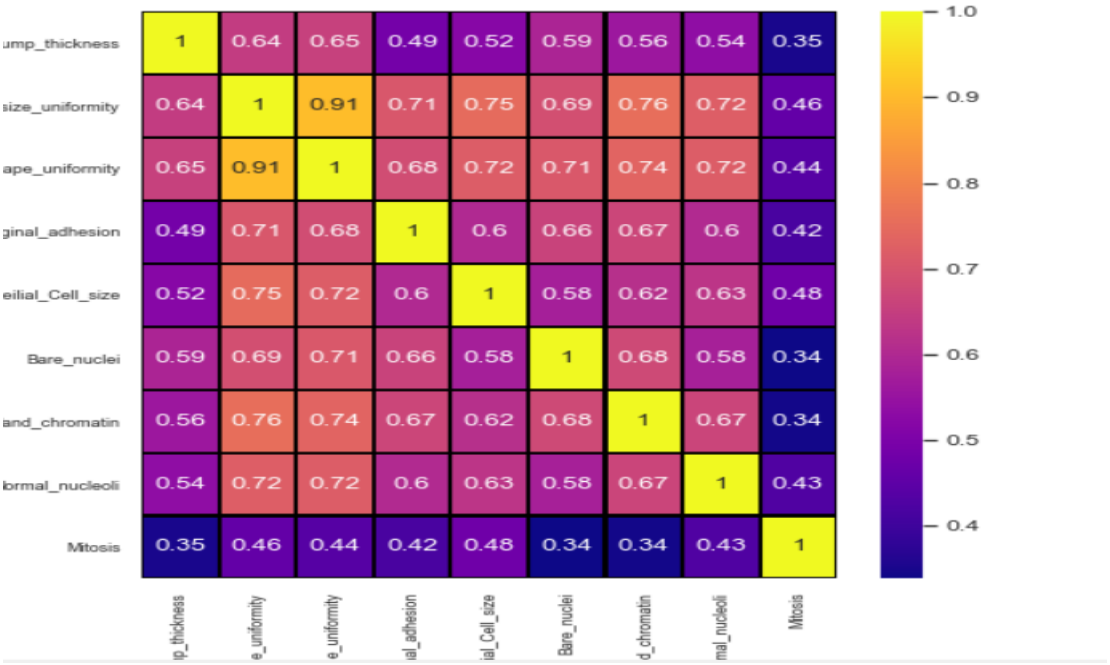

Figure 2: Heatmap based data correlation between the attributes based on SVM

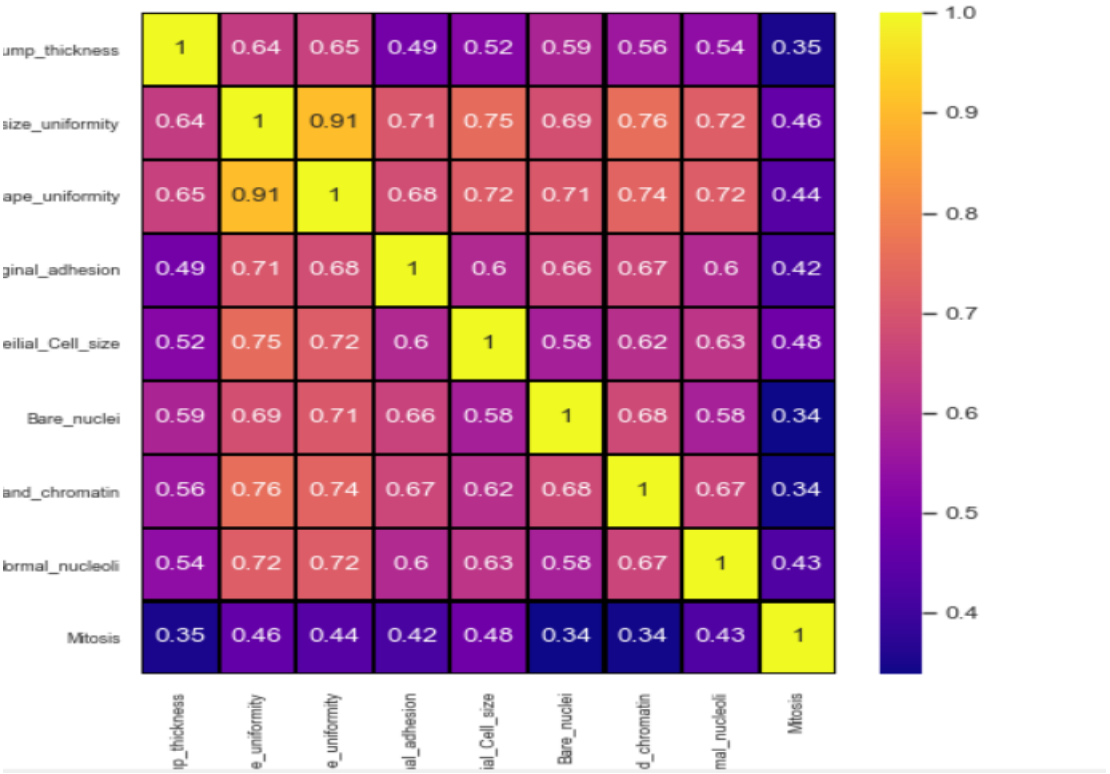

Figure 3: Heatmap based data correlation between the attributes based on ACO-RF

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

Figure 4 shows the precision, recall, f1-score, accuracy based on class labels clusters C1 and C2 in case of SVM. It clearly shows that the SVM performs better classification and it provides better results in terms of different accuracy measures. Figure 5 shows the error rate comparison based on RMSE, MSE and MAE.Figure 6 shows the average accuracy based on precision, recall, f1-score, accuracy based on class labels C1 and C2 in case of ACO-RF

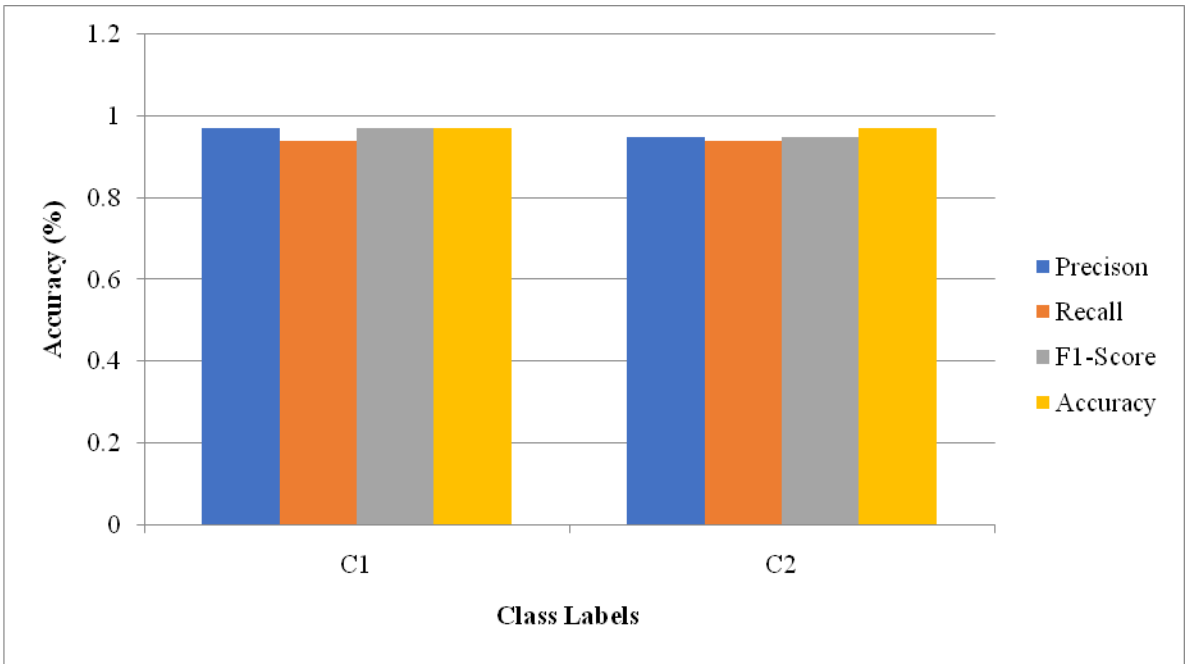

Figure 4: Precision, recall, f1-score, accuracy based on class labels C1 and C2 in case of SVM

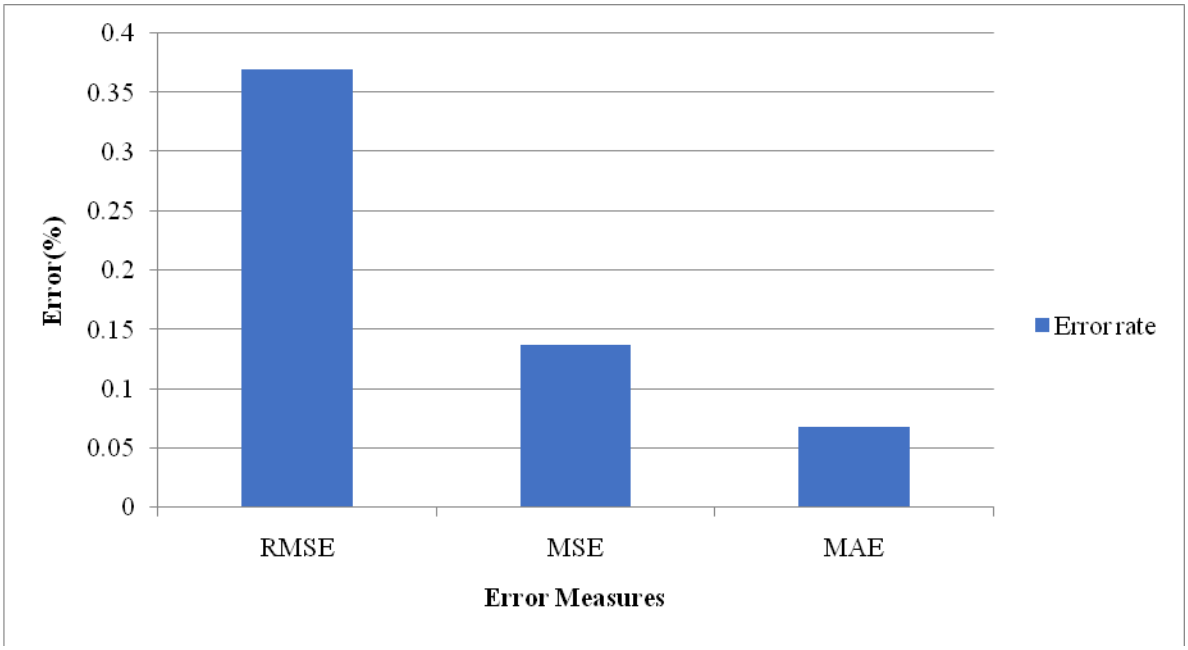

Figure 5: Error rate comparison based on RMSE, MSE and MAE

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

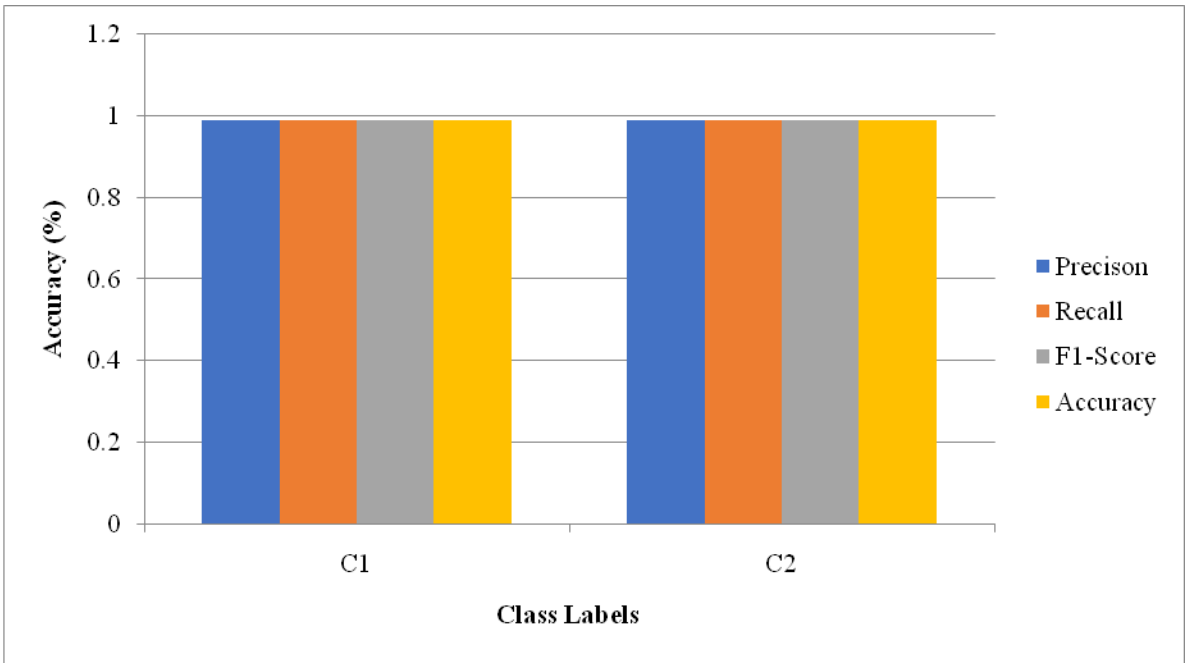

Figure 6: Average accuracy based on precision, recall, f1-score, accuracy based on class labels C1 and C2 in case of ACO-RF

4. Conclusion

This paper provides a hybridization of classification and feature selection. For the purpose of classification only, SVM algorithm has been used. It is found to be helpful in the classification based on the class labels clusters.ACO-RF combination was used for the feature selection and classification both. The results clearly indicate the applicability of the approach based on different accuracy measures like accuracy, sensitivity, precision, MAE, RMSE and MSE. It also indicates the attribute correlation and mapping based on heatmap.

References

1. Dubey AK, Kumar A, Agrawal R. An efficient ACO-PSO-based framework for data classification and preprocessing in big data. Evolutionary Intelligence. 2020 Sep 9:1-4.
2. Dubey AK, Gupta U, Jain S. Computational Measure of Cancer Using Data Mining and Optimization. InInternational Conference on Sustainable Communication Networks and Application 2019 Jul 30 (pp. 626-632). Springer, Cham.
3. Agarwal R, Srikant R. Fast algorithms for mining association rules. InProc. of the 20th VLDB Conference 1994 Sep 12 (pp. 487-499).
4. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. ACM sigmod record. 2000 May 16;29(2):1-2.
5. Jamil A, Salam A, Amin F. Performance evaluation of top-k sequential mining methods on synthetic and real datasets. International Journal of Advanced Computer Research. 2017 Sep 1;7(32):176.
6. Kumari I, Sharma V. A review for the efficient clustering based on distance and the calculation of centroid. International Journal of Advanced Technology and Engineering Exploration. 2020 Feb 1;7(63):48-52.
7. Omollo R, Alago S. Data modeling techniques used for big data in enterprise networks. International Journal of Advanced Technology and Engineering Exploration. 2020 Apr 1;7(65):79-92.

8. Kumari I, Sharma V. An efficient ICKM approach for similarity measurement and distance estimation based on k-means. International Journal of Advanced Technology and Engineering Exploration. 2020 Mar 1;7(64):73-8.
9. Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for Mining Heterogeneous Data with dynamic support. In2012 CSI Sixth International Conference on Software Engineering (CONSEG) 2012 Sep 5 (pp. 1-6). IEEE.
10. Chahar R, Kaur D. A systematic review of the machine learning algorithms for the computational analysis in different domains. International Journal of Advanced Technology and Engineering Exploration. 2020; 7 (71): 147-16.
11. Vityaev EE, Kovalerchuk BY. Relational methodology for data mining and knowledge discovery. Intelligent Data Analysis. 2008 Jan 1;12(2):189-210.
12. Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International journal of computer assisted radiology and surgery. 2016 Nov 1;11(11):2033-47.
13. Dubey AK, Gupta U, Jain S. Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science, Engineering and Information Technology. 2018 Jan;8(1):18-29.
14. Divya V, Devi KN. An Efficient Approach to Determine Number of Clusters Using Principal Component Analysis. In2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) 2018 Mar 1 (pp. 1-6). IEEE.
15. Huan Z, Pengzhou Z, Zeyang G. K-means text dynamic clustering algorithm based on KL divergence. In2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) 2018 Jun 6 (pp. 659-663). IEEE.
16. Singh R, Li K, Principe JC. Nearest-Instance-Centroid-Estimation Linear Discriminant Analysis (Nice Lda). In2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018 Apr 15 (pp. 2846-2850). IEEE.
17. López-Rubio E, Palomo EJ, Ortega-Zamorano F. Unsupervised learning by cluster quality optimization. Information Sciences. 2018 Apr 1;436:31-55.
18. Kushwaha N, Pant M, Kant S, Jain VK. Magnetic optimization algorithm for data clustering. Pattern Recognition Letters. 2018 Nov 1;115:59-65.
19. Rong Y. Staged text clustering algorithm based on K-means and hierarchical agglomeration clustering. In2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) 2020 Jun 27 (pp. 124-127). IEEE.
20. Shuai Y, Jiang C, Su X, Yuan C, Huang X. A Hybrid Clustering Model for Analyzing COVID-19 National Prevention and Control Strategy. In2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE) 2020 Jul 17 (pp. 68-71). IEEE.
21. Dubey AK, Narang S, Kumar A, Sasubilli SM, Díaz VG. Performance Estimation of Machine Learning Algorithms in the Factor Analysis of COVID-19 Dataset. Computers, Materials & Continua.2021: 66(2): 1921-1936.
22. Sebastian A, Cistulli PA, Cohen G, de Chazal P. Characterisation of Upper Airway Collapse in OSA Patients Using Snore Signals: A Cluster Analysis Approach. In2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 2020 Jul 20 (pp. 5124-5127). IEEE.
23. Alkhafaji MJ, Aljuboori AF, Ibrahim AA. Clean medical data and predict heart disease. In2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) 2020 Jun 26 (pp. 1-7). IEEE.
24. Shuai Y, Jiang C, Su X, Yuan C, Huang X. A Hybrid Clustering Model for Analyzing COVID-19 National Prevention and Control Strategy. In2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE) 2020 Jul 17 (pp. 68-71). IEEE.
25. Wang H, Dong Q, Qiang N, Zhang X, Liu T, Ge B. Task fMRI Guided Fiber Clustering via a Deep Clustering Method. In2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) 2020 Apr 3 (pp. 1420-1423). IEEE.
26. Bu F, Hu C, Zhang Q, Bai C, Yang LT and Baker T. A Cloud-Edge-aided Incremental High-order Possibilistic c-Means Algorithm for Medical Data Clustering. IEEE Transactions on Fuzzy Systems. 2020.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9830 - 9839
Received 05 March 2021; Accepted 01 April 2021.

27. Chebanenko E, Denisova L, Serobabov A. Intelligent Processing of Medical Information for Application in the Expert system. In2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) 2020 May 14 (pp. 0085-0088). IEEE.
28. Doroshenko A. Analysis of the Distribution of COVID-19 in Italy Using Clustering Algorithms. In2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP) 2020 Aug 21 (pp. 325-328). IEEE.
29. Kubo Y, Nii M, Muto T, Tanaka H, Inui H, Yagi N, Nobuhara K, Kobashi S. Artificial humeral head modeling using Kmeans++ clustering and PCA. In2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech) 2020 Mar 10 (pp. 5-7). IEEE.
30. Meniailov I, Chumachenko D, Bazilevych K. Determination of Heart Disease Based on Analysis of Patient Statistics using the Fuzzy C-means Clustering Algorithm. In2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP) 2020 Aug 21 (pp. 333-336). IEEE.
31. Roy S, Bhowmik A, Ghosh S, Roy B. Predictive Analysis & Classification of Diseases in Organs using Modified-C Clustering Technique. In2020 7th International Conference on Computing for Sustainable Global Development (INDIACom) 2020 Mar 12 (pp. 223-228). IEEE.
32. Shen H, Duan Z. Application research of Clustering algorithm based on K-means in data mining. In2020 International Conference on Computer Information and Big Data Applications (CIBDA) 2020 Apr 17 (pp. 66-69). IEEE.
33. K.Vengatesan, Dr.Selvarajan, Published a paper on "Maximize Pair Genes from Microarray using the Enhanced Fuzzy Clustering Algorithm" Journal of Pure and Applied Microbiology, Nov 2015. Vol. 9, pp. 611-618
34. Sudhagar D, Renjith JA. An Approach on Efficient Clustering Technique of High Dimensional Records. In2020 5th International Conference on Communication and Electronics Systems (ICCES) 2020 Jun 10 (pp. 860-865). IEEE.