Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DIMENSIONALITY REDUCTION BASED ON ENSEMBLE CLUSTERING APPROACH

SWETHA T[1], SAKTHIVEL G[1†] AND SANDHYA RANI D[2†]

[1]Department of Electronics and Instrumentation Engineering, Annamalai University, Annamalainager, 608002, Tamilnadu, India.
[2]Department of Computer Science and Engineering, CVR College of Engineering, Ibrahimpatnam, Hyderabad, 501515, Telangana, India.

Contributing authors: tungaturthiswetha1234@gmail.com; gsauei@gmail.com; d.sandhyarani@cvr.ac.in;
[†]These authors contributed equally to this work.

Abstract

The dimensionality reduction is still a difficult problem due to the vast volumes of data generated in social networks and bio- informatics. Here, a new dimensionality reduction approach called CCBFS-FR is devised with the inspiration of CCBFS method. This algorithm leverages genetic algorithm and is based on ensemble clustering or ensemble clustering concept. Initial population of genetic algorithm is generated using five feature rankers. A popular dissimilarity measure called Symmetric Distance Difference (SDD) is used as fitness function as similar to CCBFS. As genetic algorithm uses some stopping criteria to limit the number of populations, we use stopping criteria similar to CCBFS which was proved to be good one. The time complexity of CCBFS-FR is also less compared to CCBFS. The technique is tested on most challenging benchmark datasets with dimensionality ranging from low to medium using classification as validation. Except for Breast cancer data set we achieve good accuracies and less number of features for remaining data sets compared to literature. For Wine data set we achieve the accuracy as 96.28% with 7 features, PIMA achieved 80.35% with 4 features and Wine quality white achieved 68.72 with 7 features.

Keywords: *Dimensionality Reduction, Ensemble Clustering, Genetic Algorithms.*

1 INTRODUCTION

In this AI era, it is very advantageous for data professionals to use dimensionality reduction while working with massive data sets, complex data and performing data visualization. It facilitates data compression, which lowers calculation times and the amount of storage space needed for the data. There is an explosion of data in terms of quantity as well as quality. Generally, as the number of features increases, high dimensional data sets with a fixed amount of data points or instances get more "sparse". This problem is called "Curse of dimensionality"[1].

1.1 Necessity of feature reduction

Feature reduction, also known as feature selection, is the process of lowering the total number of features taken into account. The process of extracting pertinent features from the provided data is known as feature extraction. Approaches to feature selection look for a feature subset that best summarizes the data [2].

The features or dimensions in the dataset may accompany with few problems as follows:

1. Most of the dimensions may be non-relevant.
2. Most of the dimensions may be redundant with other dimension.
3. Classification becomes easy by removing non-relevant and duplicate dimensions.

Eliminating such unimportant, noisy, and redundant features is therefore crucial. The most popular techniques for reducing features are non-linear dimensionality reduction, principal component analysis(PCA), wavelet analysis, random projection, etc. These methods are categorized under feature extraction, where a

new set of features will be constructed from existing feature set. Another parallel technique called feature selection which selects a subset of features from the whole set. We use feature selection method in our proposed algorithm.

1.2 Motivation

A data set's "best" collection of features that presents and explains the data well is its intrinsic dimensionality. By examining these best features which effectively 'clusters' the data, one can determine the subset of features that best discriminates the data. However, the data is clustered differently by several clustering techniques. Therefore, a comparable feature subset that finds the ensemble among these several clusterings may be referred to as the "optimal feature subset".

Recently, an algorithm called as CCBFS[3] has been developed based on this idea. They developed a genetic algorithm in which initial population is selected randomly and each chromosome represents a feature subset. But as the initial population is randomly selected, the solution can be found in local optimal and each time algorithm gives different feature subsets as final output. Our method instead initializes population based on feature ranking method and it does not requires multiple runs.

2 ENSEMBLE CLUSTERING

The goal of clustering is to divide the data objects into groups with a high intra-cluster similarity and a low inter-cluster similarity [4]. However, there are other clustering techniques with varying benefits and drawbacks that may be found in the literature. Then, selecting the optimal clustering algorithm is challenging. Ensemble clustering, also known as cluster ensembling or aggre- gation of clustering [5], is a technique that can be used to solve such problems. Selecting the optimal clustering that agrees with numerous input clusterings is the fundamental notion behind ensemble clustering. In order to build a median partition, this uses as input partitionings produced by executing different clustering or partitioning algorithms or repeatedly executing the same clustering method while modifying the initial input parameters (in this case, the K-value in K-means).

The remainder of the document is structured as follows:

A few common techniques found in the literature are described in Section 3. Our suggested new dimensionality reduction algorithm is shown in Section 4. Section 5 reports the findings of the experiment. The results and future scope are summed up in Section 6.

3 METHODS USED FOR DIMENSIONALITY REDUCTION

Principal Components Analysis (PCA), the Fourier transform, heuristic approaches, and others are some of the most widely used dimensionality reduction techniques. Bayesian networks [6], selective bayesian networks [7], SBPCA [10], FRNN-FS [9], NB-tree [8], and others are further techniques for dimensionality reduction that have been reported in recent research.

3.1 Classical Techniques used to Analyze High Dimensional Data

3.1.1 Principal Component Analysis(PCA)

Pearson[11] introduced Principal Component Analysis (PCA), one of the most widely used linear methods for dimensionality reduction. Finding an orthogonal projection of high dimensional data in a low-dimensional linear subspace where the projected data's variance is highest is the goal of this technique. Although PCA has many limitations as follows, it may be successful in decreasing the number of dimensions.

- Interpretation of new features become difficult.
- Because different qualities have varying connections, it might not always be able to eliminate several dimensions without losing crucial information.
- Eigen vector calculation is not possible for really high featured data.

3.1.2 Selective Bayesian Network(SBC)

When features are redundant or irrelevant, Ratanamahatana et al.[7] employed this classifier to improve the na¨ıve bayesian classifier's performance. They gen- erated a decision tree using C4.5 initially, and they determined that the most crucial traits were those found in the first three tiers of the decision tree. They completed the union of all selected

qualities during each of the five repetitions. To check accuracy, a naive Bayesian classifier is employed.

3.1.3 Bayesian Networks

A data reduction method is proposed by Petri et al. [6] to view high- dimensional data. To transform high-featured vectors into low-featured vectors, they employ multidimensional scaling. They define an unsupervised measure called as bayesian distance, to determine the similarity between items. The naive bayesian classifier (NBC) and the class-color clarity test were used to validate the 2D data presentation. However, this test did not work with data sets that do not perform well for naive bayesian classifiers.

3.1.4 NBTree

To get around the drawbacks of using each of them alone, a combination of decision tree induction and naive bayesian classifier is employed. In order to increase the accuracy of naive bayesian classifiers in high-dimensional datasets, Kohavi[8] proposes the NBTree technique. The results demonstrated that the NBTree algorithm performs better than the C4.5 and Naive Bayesian (NB) decision tree induction techniques.

3.1.5 SBPCA

In [10], Acharya developed Supervised Bergman PCA (SBPCA), a method for reducing dimensionality that directly maximizes the prediction target in reduced dimensions. Moreover, it is demonstrated that SBPCA performs better than PCA if the classes can be divided linearly.

3.1.6 FRNN-FS

A feature selection approach based on a fuzzy rough neural networks (FRNN) was developed by Zhao et al. [9]. FRNN is used in conjunction with a heuristic backward search approach to choose features. It combines feature selection with neural networks. This algorithm has O(nlogn) time complexity.

3.1.7 Recent literature work on feature subset selection

In order to enhance the accuracy of classification of hybrid approaches with minimal evaluation consumption, Jian et al,.[22] suggests a subset of features selection strategy. In order to exploit the interactions between all the characteristics, he suggested algorithm that discards features based on their rank in the block level and also assesses all the other features that have been confirmed to be useless. A wrapper-based decision is also intended to eliminate the superfluous elements even more. The method chooses a feature subset with outstanding classification performance on all indices while using fewer evaluations, according to the theoretical evaluation and experimental results.

The paper [23] gives the latest developments in incremental feature selection methods. In addition to some unsupervised as well as supervised learning problems where the use of FSS is intriguing, various methodologies are presented, including using information theory, employing rough set-based FSS, and progressively updating feature weights.

Sweidan et al [24] proposed an enhanced correlation-based feature selection technique that maximizes the selection of non-redundant relevant feature subsets is presented in this study. Additionally, the algorithm reliably finds a small feature set, producing a statistically significant running time and noteworthy prediction performance on test data that has not before been seen. These findings demonstrate the efficiency and promise of improving feature selection procedures for a range of applications.

The study in paper [25] presents a feature subset selection method that uses the NSGA-II algorithm to generate Pareto-optimal solutions while improving accuracy and fairness objectives. They employed the F1-Score as a model performance metric and statistical discrepancy as a fairness metric.

The effective hybrid feature subset selection method known as CAPPER is developed by Thomas et al [26],. It combines feature subsets from the Wrapper and Correlation-based feature selection (CFS) approaches. The machine learning technique is then used to train the condensed, highly merited, and accurate set of features that were acquired using the CAPPER approach. The collection of attributes is then assessed by cross-validation. On several dataset domains, it was found that the CAPPER strategy outperforms the wrapper and CFS approaches.

3.2 Related work of Ensemble Clustering

According to Monti et al [12]., clustering algorithms are essentially sensitive to the initial clustering settings and the similarity measures used. Ensemble clustering is a concept that has been put out to overcome these problems. Ensemble clustering uses a variety of clusterings produced by executing different clustering algorithms or repeatedly running the same algorithm with different beginning input parameters to produce a median partition as input [5]. In order to address the issue of sensitivity to starting parameter values, this technique can also be used to express the ensemble over numerous executions of the partitioning algorithm using randomly assigned re-start points. Finding ensemble amongst the exponentially many possible input clusterings is an NP-hard task. As a result, numerous heuristics are put out in the literature[13] to deal with this issue.

3.2.1 Ensemble Clustering Approximation Algorithms

A number of algorithms are suggested to get a ensemble over the groupings. There are essentially two methods used to accomplish this. The first option is to use some dissimilarity measures to select one of the clusterings as the ensemble. The second step is to rearrange clusters in order to reach an agree- ment. A wide range of approximation algorithms, including graph partitioning algorithms such as CSPA, HGPA, MCLA[14], ensemble functions[15],and majority rule (MR) best-of-k (BOK), best one element move (BOEM), average link, simulated annealing, etc.[13], are available in the literature. Dis- similarity matrices are used by all approximation techniques in the literature to determine ensemble. However, these approximation methods, aside from BOK, have time complexity of at least $O(n^2)$. Additionally, BOK has an $O(k^2n)$ linear time complexity, where 'k' is the number of input clusterings.

3.2.2 Dissimilarity Measure

A number of metrics, including the dissimilarity measure[13][16], Quality Par- tition Index (QPI)[17], and Normalized Mutual Information (NMI)[14], can be used to calculate the distance between two clusterings, X_1 and X_2. Dissimilarity measure, which is defined as follows, has been utilized in the majority of works.

$$d(X1, X2) = (b + c) \; or \quad \binom{n}{2} - (a + b) \tag{1}$$

Where
a is the number of object pairings clustered in X_1 and X_2.
b is the number of object pairs that are clustered in X_1 but not in X_2.
c is the number of object pairs that are clustered in X_2 but not in X_1.
d is the total number of object pairs in X_1 and X_2 that are not clustered.

The clustering that differs the least from the other clusterings is the best clustering. Best-of-k(BOK) approximation approach is employed here as the time complexity is linear $O(k^2n)$[13], where k' is the number of input cluster- ings and n' is the number of cases. Out of the k input clusterings, it selects the best clustering. The algorithm in question is a 2-approximation [18]. Because the K-means approach is so straightforward, it is utilized here to create input clusterings.

4 PROPOSED ALGORITHM

In this paper a novel ensemble clustering based feature selection algorithm using genetic algorithm approach is proposed. General steps in genetic algorithms are:

1. Fix the size of the population
2. Generate the chromosomes in the first population using feature ranking algorithms.
3. Using fitness criteria select the best chromosome from initial population and retain the same in next population
4. In the upcoming population generate the rest of the chromosomes by applying 'mutation' and 'crossover' functions on previous population chromosomes.
5. Repeat step 3 and 4 until the best chromosome is selected based on some fitness function.

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

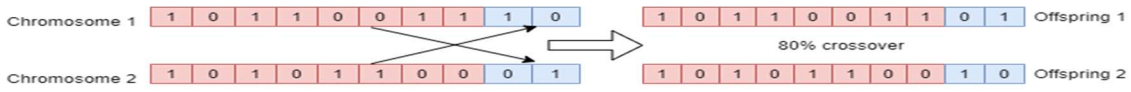

Fig. 1 Crossover in genetic algorithm

4.1 Feature Selection using Genetic Algorithm CCBFS-FR

Genetic algorithm and its functions have been used to generate various feature subsets. In the first step chromosomes are selected by running most popular feature ranking algorithms like Symmetric uncertainty, information gain, gain ratio, relief and OneR. Here 'Bit 1' positions of chromosome constitutes feature subset. Figure1 describes the method of CCBFS-FR.

As a fitness function, the dissimilarity value mentioned in equation (1) from BoK is employed. Using a genetic algorithm, this method selects the 'selection' approach, which allows the best chromosomes to be kept for the following population. The resulting solution can become locally optimal by using this 'selection'. The two genetic operators that are utilized to generate the next population are 'crossover' and 'mutation'. When the optimal chromosome is chosen from the preceding popu- lation, the algorithm comes to an end. With n' instances, k' input clusterings, and K' clusters, our approach requires O(Knk) time. Algorithm 1 describes the CCBFS-FR technique.

It is proved in the paper [3], [19] that ensemble clustering can be used efficiently for feature selection problem. They have constructed one synthetic dataset with four clear and separable clusters with two independent dimensions x, y. Then, increased the number of dimensions to 6 with the combination of x,y. By using a brute force approach on the synthetic dataset, they proved that ensemble clustering will pick the independent features. This is reason why we have selected ensemble clustering for our experiment.

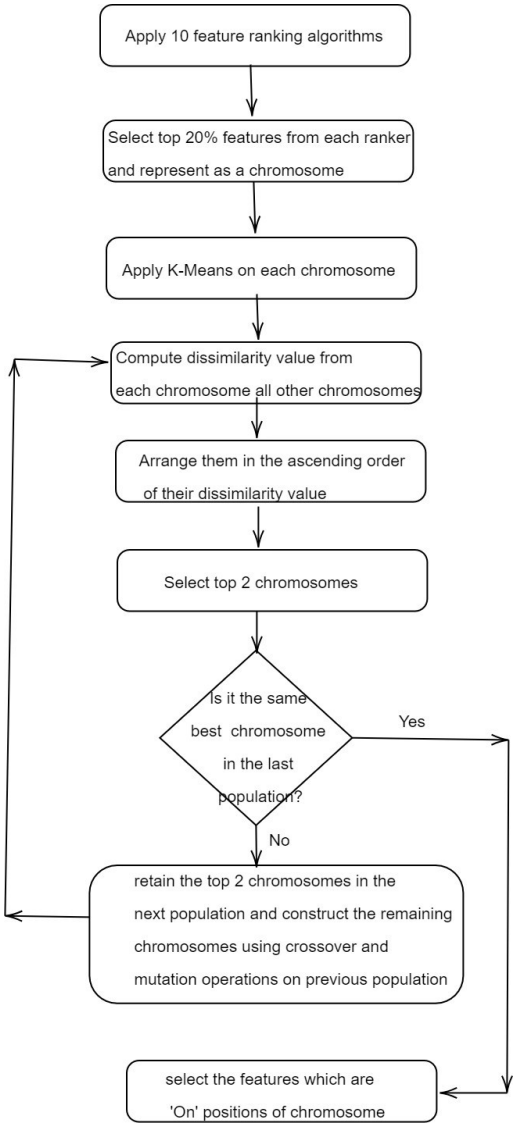

Fig. 2 Flow diagram of proposed method

5 Experiments and Results

5.1 Data Sets

In this paper, the work is done with UCI machine learning datasets including PIMA, breast cancer (BC), and wine quality-white (WQW), wine. PIMA and BC are binary class datasets where as WQW and Wine are multiclass datasets. The literature reports good classifier accuracies for data sets wine and breast cancer. PIMA data set, along with wine-quality (white), are not properly clas- sified in the literature. These data sets are chosen so as to test our approach in scenarios when the accuracies of the classifier are both good and poor. These datasets are challenging datasets in the literature. The quantity of features and other

related information about these data sets are displayed in Table 1. These data sets are more challenging datasets from literature. Though the number of features are ranging from 8 to 13, the instances are varying from hundreds to thousands. We are interested to see how the results will vary with the increase in number of instances. PIMA, Wine and breast cancer are the data sets used very frequently in machine learning experiments and there were no results available so far for Wine quality dataset in the literature. To work more on such data sets, we have chosen PIMA, wine, wine quality white and breast cancer data sets.

Algorithm 1 CCBFS-FR Algorithm

Input: Dataset is D. N is the data set size and M is the feature set, where F={f1, f2, f3, . . . , fM }, W={w1, w2, w3, . . . , wM },
FR={*Gain ratio, Information gain, OneR, ReleifF and Symmetric Uncer- tainty*}
▷ FR represents Feature Ranking Algorithm
Output: FS: Feature subset
1: FS ← NULL
2: **for each** x ← 1 to n in FR **do** ▷ Apply feature ranking algorithm on D
3: (F, W)x = Feature Ranking(F,x)
4: **endfor**
5: Apply K-Means clustering algorithm on the output of each ranking by using feature weight as dimension.
6: Select features from the cluster that is having all top most features that has highest weights.
7: Use these feature subsets as chromosomes in first population.
8: Chromosome := bitvector with length equal to total number of features in the dataset. bit '1' means feature being selected and bit '0' means feature not selected.
9: Apply K-means on each chromosome while selecting 'K' using Dunn index
10: Apply best-of-k(BoK) to choose best chromosome by using SDD(Symmetric Dissimilarity Distance) as fitness function
11: Choose top two chromosomes from that population and retain them in next population.
12: Generate remaining chromosomes using 'Crossover' and 'Mutation' operations.
13: Repeat this until the top chromosome is selected as the best one similar to the previous population.
14: FS ← features from best chromosomes.
15: **return** FS
▷ Final feature subset.

Table 1 Details of UCI datasets

Dataset	# Instances	# Features	# Classes
PIMA	768	8	2
BC	699	10	2
WQW	4878	11	11
Wine	178	13	3

5.2 Experiments and results obtained with CCBFS-FR

As reported in the Table 2, CCBFS-FR is outperforming with CCBFS and other methods mentioned in the literature. In PIMA, Breast cancer and Wine dataset CCBFS-FR has got same number of features as CCBFS and on par with other methods and number of features in WQW is less compared to CCBFS method. It is clear from the results that except for breast cancer data set, our method is selecting less number of features with better accuracy compare SBC method in literature. But the difference is not significant. Accuracies achieved by using J48 classifier is higher in all datasets for CCBFS-FR method compared to literature. Thus it can be said that CCBFS-FR is more powerful method compared to CCBFS and other methods. Further, the running time of both the methods is compared and results shows that CCBFS-FR is taking less time compared to CCBFS and also number of generations are less in CCBFS-FR. Table 3 gives the comparison of running time of both the methods.

Table 2 Result Analysis Of CCBFS-FR With CCBFS And Other Methods From Literature

Data set	CCBFS-FR		CCBFS		Others		
	Final Feature subset	Accuracy	Final Feature subset	Accuracy	Other Methods	Final Feature subset	Accuracy
PIMA(8)	**4**	**80.35**	**4**	77.8	SBC	5 –	79 77

					SBPCA		
BC(10)	5	96.86	5	96.7	SBC SBPCA FRNN-FS	**4** – 4	**97** 96 93.8
WQW(11)	**7**	**68.72**	8	62.68	- -	–	–
Wine (13)	**7**	**96.28**	**7**	95.03	FRNN-FS	6	94.77

Results obtained using CCBFS-FR and CCBFS are compared and plotted in Figure 3.

Table 3 Comparison Of Running Time Of CCBFS-FR And CCBFS

Dataset	**Number of generations**		**Runtime(sec)**	
	CCBFS-FR	CCBFS	**CCBFS-FR**	CCBFS
PIMA	**6**	13	**187**	2098
BC	**12**	27	**548**	3452
WQW	**11**	31	**421**	3876
Wine	**21**	46	**654**	7648

To test the results obtained in CCBFS-FR and CCBFS, further features obtained in both the methods are analysed. As an example results of Wine dataset is provided in the Table 4. It is observed from the results that 5 out of 7 features are common in both the methods and feature *Flavanoids* is having high impact on the result.

Table 4 Features Obtained For Wine Data Set By CCBFS-FR Method And CCBFS Method

CCBFS-FR method	CCBFS method
total phenols alcohol proanthocyanis magnesium color intensity Hue Flavonoids	alcohol magnesium ash proanthocyanis color intensity malic acid Hue
96.28% accuracy	**95.03% accuracy**

5.3 Comparing the cluster results with cluster validity indices

The clustering method's ability to produce compact clusters is confirmed using cluster validity indices. Following are the few cluster validity indices from the literature [20]:

- Dunn Index
- Silhouatte Index
- Root-mean-square standard deviation (RMSSDT)
- Davies-Bouldin (DB)

The Dunn index, which finds compact and well-separated clusters, is one of the most often quoted indices[21]. Dunn index is used to test the outcome.

5.3.1 DunnIndex

$$\text{Dunn} = \frac{dmin}{dmax} \quad (2)$$

Where d_{min} represents the minimal distance among objects of two different clusters and d_{max} represents the maximum distance within the same cluster. The measure Dunn index value high means it is a having the better clustering quality where the clusters are well separable. Final results are cross verified with number of clusters obtained using CCBFS-FR method and *Dunn* index. Except WQW, it is proved in PIMA, BC and Wine that number of clusters are same in both the cases.

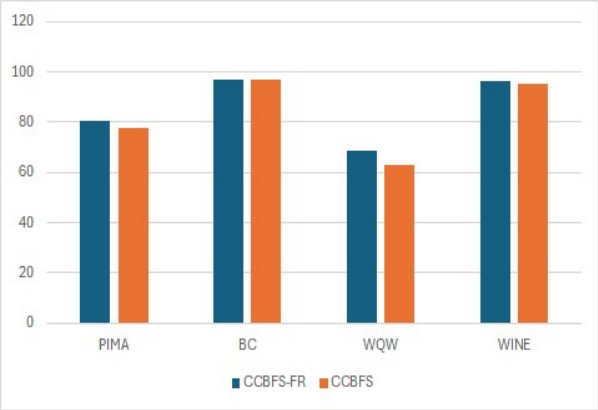

Fig. 3 Comparison Of CCBFS-FR And CCBFS

5.4 Discussion

The literature has numerous techniques for reducing dimensionality. The approach of consensus clustering for dimensionality reduction is the novel method we proposed here. It is proved that this method is simple and takes less time for clustering as well as chooses the consensus one in less time. This method is evaluated against other approaches from the literature that have made use of the same experimental data sets.

Any high-dimensional data can be reduced to two or three dimensions using the Bayesian network method. To determine the pairwise distance between two vectors, they employed the unsupervised Bayesian distance metric. It is a challenging task to estimate distribution when using an unsupervised distance measure. When building a simplified decision tree using C4.5, the SBC algorithm chooses a set of characteristics that are limited to the intial three levels. It might not always be effective to choose just three levels.

The results achieved using CCBFS-FR are higher compared to CCBFS and on par with other methods in the literature. In CCBFS-FR method also genetic algorithm is used as similar to CCBFS, but initial population was selected based on the outputs of feature ranking algorithms. Time complexity of this algorithm is O(gpn), with 'n' number of instances, 'g' is the number of generations and 'p' is number of chromosomes per population. Wine dataset is achieving 96.28% accuracy CCBFS with same number of features. Even for breast cancer, wine-quality(white) and PIMA, CCBFS-FR performed better compare to CCBFS and other methods from literature. From the experiments conducted, it is concluded that CCBFS-FR always selects less number of features than CCBFS. Also it is shown from the results that running time of CCBFS-FR is less compared to CCBFS as the number of generations in CCBFS-FR is always less. As CCBFS-FR does not choose the features randomly in the initial population, it always selects the best feature subset in local minima.

6 CONCLUSIONS

One of the hardest problems in data mining is clustering high dimensional data. The data set is divided into distinct clusters by removing unwanted dimensions. The most popular dimensionality reduction method PCA suffers from huge time complexity and does not select a subset of features from full feature set, instead a new set of features will be constructed. Thus, it can be said it is not feasible for high-dimensional data with large number of features. Inspiring from CCBFS method, this method is improvised by using feature ranking in it and named the method as CCBFS-FR method. This algorithm is tested on most challenging benchmark datasets from UCI machine learning repository. In CCBFS-FR method, genetic algorithm is used with initial population generated using feature ranking algorithms. From the results it is clear that CCBFS-FR method is performing better compared to CCBFS method and other methods from literature. The fitness function called dissimilarity measure has been used in both CCBFS-FR, CCBFS and it should be minimum always. CCBFS-FR, CCBFS both uses same stopping criteria and is also proved as good choice. Time complexity of CCBFS-FR is less compared to CCBFS but this may not scalable high dimensional datasets. In future, this method can be improved to minimize time complexity.

The summary of this paper contributions is as follows:

- Developed a novel genetic algorithm based ensemble clustering algorithm for feature subset selection.
- It finds the solution in local minima.
- Number of generations are less compared to CCBFS algorithm.

- Time complexity is low compared to CCBFS method and high as the number of generations increases with number of dimensions.
- One limitation of our algorithm is that, it is not converging at stopping criteria when the number of dimensions are more than 100. We experimented on ISOLET dataset which is having 625 features, where convergence did not take place.

Further, we would like to enhance our method to work on dimensions that are more than 100. As genetic algorithms takes more time to converge at solution due to many iterations, we would like to devise a fast algorithm which does not uses iterative method.

References

[1] Wang, W., Yang, J.: Mining high dimensional data. In: Maimon O , Rokach L (eds) Data Mining and Knowledge Discovery Handbook,Springer, pp. 793–799 (2005)

[2] Yang, J., Tiyyagura, A., Chen, F., Hanover, V.: Feature subset selec- tion for rule induction using ripper. In: Proceedings of Genetic and Evolutionary Programming, pp. 117–136 (1998)

[3] Rani, D.S., Rani, T.S., Bhavani, S.D.: Ensemble clustering for dimension- ality reduction. In: 2014 Seventh International Conference on Contempo- rary Computing (IC3), pp. 148–153 (2014). https://doi.org/10.1109/IC3. 2014.6897164

[4] Han, J., Kamber, M.: Data Mining Concepts and Techniques, Morgan Kaufmann Series in Data Management Systems, (1999)

[5] Ghosh, J., Acharya, A.: Cluster ensembles. In: WIRE's Datamining Knowledge discovery,John Wiley and Sons, vol. 1, pp. 305–315 (2011)

[6] Kontkanen, P., Lahtinen, J., Millymaki, P., Tirri, H.: Unsupervised bayesian visualization of highdimensional data. In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Min- ing(KDD), 325–329 (2000)

[7] Ratanamahatana, C.A., Gunopulos, D.: Feature selection for the naive bayesian classifier using decision trees **17**, 475–488 (2006)

[8] kohavi, R.: Scaling up accuracy of naive bayes classifier: a decision tree. Knowledge Discovery and Datamining(KDD), 202–207 (1996)

[9] Zhao, J., zhili zhang: Fuzzy rough neural network and its application to feature selection. International Journal of Fuzzy Systems **4**, 270–275 (2011)

[10] Acharya, S.: Transductive denoising and dimensionaity reduction using total bregman regression, siam, pp. 514–518 (2006)

[11] Pearson, K.: On lines and planes of closest fit to systems of points in space. In: Philosophical Magazine 2, pp. 559–572 (1901)

[12] Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Ensemble clustering: A resamplingbased method for class discovery and visualization of gene expression microarray data. Machine Learning, 91–118 (2003)

[13] Goder, A., Filkov, V.: Ensemble clustering algorithms : Comparision and refinement. In: Proc SIAM International Conference on Data Mining, pp. 109–118 (2008)

[14] Strehl, A., Ghosh, J.: Cluster ensemblesa knowledge reuse framework for combing multiple partitions. Journal of Machine Learning, 583–617 (2002)

[15] Topchy, A., Jain, A.K., Punch, W.: A mixture model for clustering ensem- bles. In: Proceedings of SIAM International Conference on Datamining, SIAM, pp. 379–390 (2004)

[16] Bertolacci, M., Wirth, A.: Are approximation algorithms for ensemble clustering worthwhile?, siam, pp. 437–442 (2007)

[17] Varin, T., Saettel, N., Villain, J., Lesnard, A., Dauphin, F., Bureau, R., Rault, S.: 3d pharmacophore, hierarchical methods, and 5ht4 receptor binding data. In: Enzyme Inhib Med Chem, pp. 593–603 (2008)

[18] Filkov, V., Skiena, S.: Integrating microarray data by ensemble cluster- ing. Journal of Artificial Intellegence Tools,WSPC, 863–880 (2004)

[19] D. Sandhya Rani, M.R.K.S.V.P.K.M.A. G. Bala Krishna: Gbjof: Gradi- ent boosting integrated with jaya algorithm to optimize the features in malware analysis. International Journal on Recent and Innovation Trends in Computing and Communication **11**, 538–548 (2023)

[20] Kovacs, Legany, F., Babos, A.: Cluster validity measurement tech- niques. In: Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence, pp. 18–19 (2005)

[21] Saitta, S., Raphael, B., Smith, I.F.C.: A bounded index for cluster validity. Lecture notes in computer science,Springer (2007)

[22] Shen, Jian and Tan, Xiaodong and Yin, Yixi: A Multi-Stage Feature Subset Selection Approach for Classification. Proceedings of the 2023 3rd International Conference on Big Data, Artificial Intelligence and Risk Management, Association for Computing Machinery. ACM digital library, 196 – 200 (2024).

[23] Villa-Blanco, C., Bielza, C. & Larrañaga, P. Feature subset selection for data and feature streams: a review. *Artif Intell Rev* **56** (Suppl 1), 1011–1062 (2023). https://doi.org/10.1007/s10462-023-10546-9.

[24] D. Sweidan, "Correlated, Yet Independent A Correlation Based Feature Selection Algorithm," *2024 Intelligent Methods, Systems, and Applications (IMSA)*, Giza, Egypt, IEEE 2024, pp. 530-537, doi: 10.1109/IMSA61967.2024.10652865.

[25] Ayaz Ur Rehman, Anas Nadeem, Muhammad Zubair Malik, Fair Feature Subset Selection using Multiobjective Genetic Algorithm, GECCO '22 Companion, ACM. https://doi.org/10.1145/3520304.3529061

[26] N, Thomas Rincy and Gupta, Roopam, An efficient feature subset selection approach for machine learning, Kluwer Academic Publishers, 2021. https://doi.org/10.1007/s11042-020-10011-7