CLUSTER VALIDATION OF EVOLUTIONARY CLUSTERING ALGORITHM FOR MULTIVARIATE DATASETS

Ms. Jyoti Lakhani, Dr. Ajay Khunteta, Dr. Anupama Chowdhary, Dr. Dharmesh Harwani

Abstract - Clustering is an unsupervised technique that is used when the information about data is inadequate. Clustering algorithms identify clusters in the dataset and acquire information on the number of clusters to be identified as a prior parameter. In the present work, an adaptive evolutionary clustering algorithm (AECRAM), which evolves automatically, was tested to identify the number of clusters in a multivariate dataset. It is important to test whether the identified clusters are actual clusters or not, and if difference exists in them, then to find out these differences, relative to their actual clusters using cluster validity indexes. Using adaptive evolutionary algorithm cluster validation was done in this work and compared with K-means and EKM algorithms for their relative efficiency. The observations made in the study clearly indicated that the proposed adaptive clustering algorithm could detect correct and accurate number of clusters in the multivariate datasets. The quality of the clusters identified by the AECRAM has been found to be superior to the other two algorithms used. The present study is also useful in other research where relative cluster validations among adaptive evolutionary clustering algorithms are required.

Keywords: Cluster, Validity Index, Dunn Index, Malignant, Benign.

I. INTRODUCTION

Cluster analysis is an unsupervised technique used for the data analysis in cases where very less information is available about the data. In cluster analysis, one of the important parameter is to find out the number of clusters that is done by a clustering algorithm. Once clusters are identified, next step is to validate the number and quality of the clusters identified. Cluster validation also includes comparison of overall results executed by the selected clustering algorithms. In other words, cluster validity is a measure of goodness of a clustering algorithm in terms of number and quality of the identified clusters using a set of parameters. There are variety of cluster validity indexes available in literature (Halkidi et al. 2001), (Wu et al. 2009), (Milligan and Cooper 1985), (Dimitriadou et al. 2002), (Maulik and Bandyopadhyay 2002), (Pal and Bezdek 1995), (Wang et al. 2009), (Rendon et al. 2011). According to Theodoridis and Koutroumbas (2008) there are three common approaches used for the cluster validation namely internal cluster validation, external cluster validation and relative cluster validation. The internal cluster validation is based on the datasets, cluster organization and intrinsic information available in the data. The external cluster validation is based on the prior knowledge that is available about the data in the form of predefined classes. The third cluster validation technique is the relative cluster validation in which results of clustering algorithm are compared and analyzed for the best results by varying parameters. A relative method is also used to compare and identify the best clustering method for a specific dataset.

A comparison of the thirty internal cluster validity indexes has been presented by Milligan and Cooper (1995) for hierarchical clustering algorithms. Dimitriadou et. al. (2002) performed a comparative study of fifteen external cluster validity indexes for binary data. A systematic survey of cluster validation indexes has been also done for the post genomic data (Handl et. al., 2005). A relative study of 16 external cluster validation indexes has been presented by using k-means algorithm (Wu et al. 2009a), (Wu et al. 2009b). Moshtaghi et al. (2018) developed online version of Xie-Beni and Davies-Bouldin internal validity indices, with and without forgetting factors for data streaming. In the literature, researchers did not find any dedicated cluster validity index for adaptive evolutionary clustering algorithm. And If available, the available cluster validity index cannot be applied for clusters with arbitrary shapes. Zhu et al. (2018) introduced a novel WCH index that can effectively process the datasets with high overlapping and can accurately find the corresponding cluster validity index.

Cheng et al. (2019) proposed a novel Local Cores-based Cluster Validity (LCCV) index to improve the performance of Silhouette index. The limitation of LCCV index is that it is not yet analyzed for hierarchical clustering. Da Silva et al. (2019) extended cluster validity indices (iCVI family) to incremental cluster validity indices. The members of iCVI family are Calinski-Harabasz (iCH), I index and Pakhira-Bandyopadhyay-Maulik (iI and iPBM), Silhouette (iSIL), Negentropy Increment (iNI), Representative Cross Information Potential (irCIP), Representative Cross Entropy (irH), and Conn Index (iConn Index). These for fuzzy adaptive resonance theory (ART)-based clustering algorithms and analyzed for the detection of the correct number of clusters in high-quality partitions, under- and over-partitioning. In such cases, performance of Partition Separation (PS) index, the incremental Xie-Beni (iXB), and the incremental Davies-Bouldin (iDB) was analyzed and effective than other non-incremental indexes were confirmed. The focus of this communication is to study and analyze cluster validation indexes for the proposed adaptive evolutionary clustering algorithm to evaluate the suitability and to provide accuracy. The present study is also useful in other research where relative cluster validation among adaptive evolutionary clustering algorithm is required.

II. MATERIAL AND METHODS

a. **Evolutionary Clustering Algorithm:** An adaptive evolutionary clustering algorithm Auto-Evolving Clusters based on Rejection and Migration (AECRAM) has been proposed already (Lakhani et. al., 2016). The algorithm is able to cluster a given item in the datasets automatically into natural clusters and importantly, the information on number of clusters a-priori is not required. The algorithm considers all items in a dataset first as a single cluster.

Two operators namely separation and migration are used during the cluster evolution to auto evolve the cluster and to find out the remaining clusters in the dataset. The partitioning of clusters continues until the total individual items get adapt to their clusters and subsequently the population get stable. There are some inputs required for the algorithm that can be described stepwise as 1. Un-clustered population P of N number of individuals, 2. Threshold for health of the cluster, 3. Distance measure-Euclidian/Maratthan, 4. Threshold for maximum distance of individual data items from its centroid and 5. Number of maximum migratory individual items at a time.

The algorithm is executed and as an output, a well clustered population with N number of individuals, all adapted to its cluster is provided. The default values for the parameters for the proposed adaptive evolutionary algorithm are: Threshold for health >=1; Distance measure= Euclidian Distance; Threshold for maximum distance >=2; Maximum number of iterations/generations =50; Population size = as per the dataset; Recombination= single point; Mutation – single bit, Selection criteria = all (by default). The cluster validity indices were used in the present research to validate the numbers of clusters as identified by the proposed evolutionary algorithm.

b. **Datasets:** Four multivariate datasets were downloaded from UCI Machine Learning Repository (Dua and Karra, 2017) for the present study. The details of these datasets have been furnished in Table 1. The data was processed, discretized and scaling was performed. The class attributes were removed from the datasets 1, 2 and 4 before proceeding further.

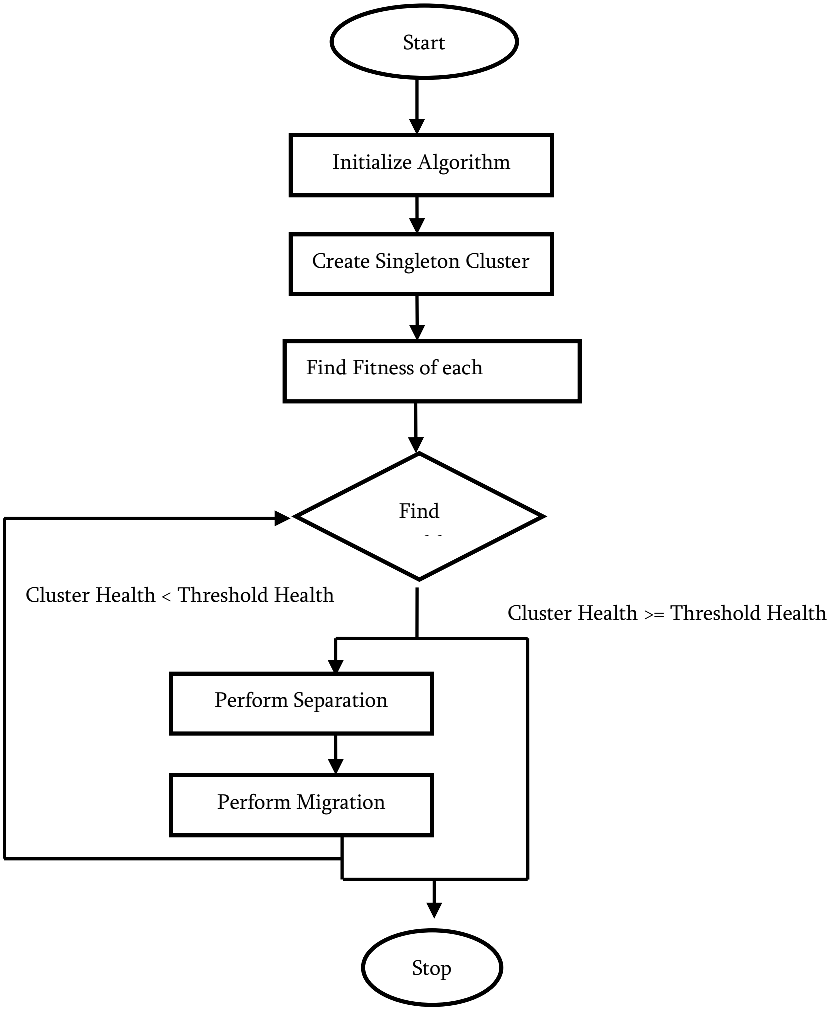

Figure 1: Execution of the adaptive evolutionary clustering algorithm (AECRAM)

c. **Cluster Validation:** For cluster validation, the three methods as described by (Theodoridis and Koutroumbas) 2008) were used. NBClust method of R package was used that uses 30 numbers of cluster validity indexes. Two statistical measures standard Deviation, F1-measure and two external cluster validity measures Silhouette index, Cluster Validation Indices (CVIs), and Dunn index were used for explicit study.

Table 1: Datasets used in this study and their properties

Datasets	Name of the Datasets	Properties
Dataset1	Iris	Dataset - Multivariate Attribute-Real Instances- 150 Attributes- 4
Dataset2	Wine	Dataset - Multivariate Attribute-Integer, Real Instances- 178 Attributes- 13
Dataset3	Contraceptive Method Choice (CMC)	Dataset - Multivariate Attribute-Categorical, Integer Instances- 1473 Attributes- 9
Dataset4	Breast Cancer Wisconsin (Diagnostic)	Dataset - Multivariate Attribute-Real Instances- 569 Attributes- 32

The proximity metrics of the clusters were developed by using ACRAM, K-means and EKM algorithms for three datasets and standard deviation was calculated by tracing the samples clustered correctly and incorrectly (randomly selected). F-measure was calculated to measure the accuracy of the clusters. F-measure is a harmonic mean of precision p and recall r. Further, p is the number of correct/positive results divided by the number of all positive results returned by a classifier while r is the number of correct positive results divided by the number of all relevant samples. F measure gives its best value at 100% and worst value at 0%.

$$\text{F measure} = 2^*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (1)$$

$$\text{where Precision} = TP/TP+FP \quad (2)$$

$$\text{and Recall} = TP/TP+FN \quad (3)$$

The silhouette analysis measures the quality of clusters and estimates the average distance between clusters. For each observation i, the silhouette width si was calculated by performing the following steps-

i. For each observation n, the average dissimilarity intra_diss between n and all other points of the cluster to which n belongs was calculated.

ii. For all other clusters C, to which n does not belong, the average dissimilarity d(n,C) of n to all observations of C was calculated. The smallest of these d(n,C) was defined as inter_diss=minC d(n,C). The value of inter_diss can be observed as dissimilarity between n and its "neighbor" cluster, i.e., the nearest one to which it does not belong.

iii. Finally the silhouette width of the observation n was defined by the formula: Sill(X)=(inter_diss−intra_diss)/max(intra_diss,inter_diss).

If the observations are falling within a large Si (almost 1) then Silhouette width can be interpreted as a very well clustered. A small Si (around 0) means that the observations lies between two clusters. The observations with negative Si are probably be placed in the wrong cluster.

The Dunn index clustering validation measure was computed by using following steps:

i. For each cluster, the distance between each object in the cluster and the objects in the other clusters was computed

ii. Minimum of pairwise distance as the inter-cluster separation (min.separation) was used

iii. For each cluster, the distance between the objects in the same cluster was computed

iv. Maximal intra-cluster distance (i.e maximum diameter) as the intra-cluster compactness was used

v. Dunn index (Dunn) was calculated using the following formula: Dunn=min.separation/max.diameter (4)

If the dataset contains compact and well-separated clusters, the diameter of the cluster is expected to be small and the distance between the clusters is expected to be large. Thus, Dunn index can be maximized (27).

III. RESULTS AND DISCUSSION

a. Validation of the number of clusters

Internal cluster validation was performed to know whether the number of clusters identified by the proposed adaptive evolutionary clustering algorithm (AECRAM) is correct or not. We executed the proposed AECRAM algorithm and the number of clusters, thus generated, were identified. Then the experiment in R tool (R Core Team, 2013) with NbClust package (Charrad et. al, 2014) was performed to identify the number of clusters in the four datasets under investigation (Table 2).

Table 2: Number of clusters identified

Dataset	**No. of clusters identified**	
	NBClust in R	AECRAM
Dataset 1	3	3
Dataset 2	2	2
Dataset 3	3	3
Dataset 4	2	4

The NbClust of R package and adaptive evolutionary algorithm AECRAM identified the same number of clusters except for the dataset 4 (Table 2). These results were further analyzed to crosscheck mismatches, if any. In external cluster validation approach the results of the cluster analysis were compared with the external results. This is performed with the datasets for which the class labels were already known. As the true cluster members were already known, the approach was mainly used for selecting the most accurate clustering algorithm for the specific dataset. Thus, the external cluster validation was performed for all four datasets.

The optimal number of clusters was further identified by fclust package (Ferraro and Giordani, 2015) of R tool using Silhouette width index and gap statistic index. The results have been furnished in figures 2a and 2b. The fclust package also identified 4 clusters in the dataset therefore it can be stated that there are four possible clusters in the dataset4. One of the probable reasons of identifying four clusters in the dataset4 could be devoted to few extreme cases (outliers) (false positives and false negatives) or boundary values that were not considered in the original dataset.

The adaptive evolutionary clustering algorithm, AECRAM, detected false positives and false negatives as the separate clusters. To cross validate this assumption for the clusters identified by the proposed evolutionary algorithm, we cross-examined the obtained results by performing statistical analysis using Naïve-bayes classifier (Lewis 1998), (McCallum and Nigam 1998) with Weka tool (Hall et al. 2009), (Frank et al. 2016).

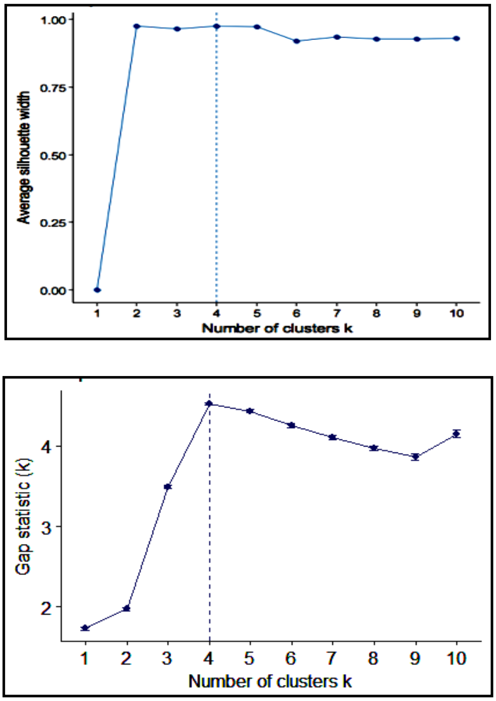

Figure 2: Execution of fclust in R package to identify the optimal number of clusters

The experiment was performed to cross check the actual classes in the dataset. As described in the Wisconsin dataset (Breast Cancer Wisconsin Dataset. Available at: UCI Machine Learning Repository) there are two breast cancer tumor classes, benign and malignant. The confusion matrix for the experiment has been furnished in figure11. Before discussing confusion matrix, it is worth to mention that the Wisconsin dataset contains records of 212 malignant and 367 benign cases (total n=569 cases/records).

However the confusion matrix created by 10 fold cross validation Naïve-bayes classifier indicated 183 cases that were actually identified as malignant whereas 25 cases were identified as benign. In addition, 332 cases were truly identified as benign and as malignant 29 cases. The confusion matrix for the dataset has been presented in table 3.

Table 3: Confusion matrix of Naïve Bayes classifier for the Wisconsin breast cancer diagnostic dataset

n=569	Malignant	Benign
Malignant	183	29
Benign	25	332

For relative cluster validation, K-means and Evolutionary K-means algorithms were implemented using Java 8 using eclipse IDE. These algorithms were compared with AECRAM for dataset1, dataset2 and dataset3 cluster validation. The relative cluster validation process was performed using two statistical measures (Standard deviation and F1-measure) and two Cluster Validation Indices (CVIs) (Silhouette index and Dunn Index).

Table 4: Cluster validation of the three clustering algorithms for dataset1 for 50 runs

Algorithms	**Standard Deviation**	**F-measure (F1Score)**	**Silhouette Coefficient**	**Dunn Index**
K-Means	4.11	89.45%	0.4599482	0.2649665
Evolutionary K-Mean(EKM)	3.94	89.98%	0.4587651	0.2645856
AECRAM	2.06	91.01%	0.4627765	0.2589768

Cluster validation of the clusters identified by K-means, evolutionary k-means and AECRAM was performed with dataset1 (Table 4), dataset2 (Table 5) and dataset3 (Table 6). By looking at the values observed with standard deviation, the algorithm AECRAM identified the clusters more accurately relative to the algorithms K-means and EKM (Tables 4). Thus identified clusters were also precise in terms of their F-measure with F1score of 91.01% whereas K-means and EKM resultant clusters resulted in F1Score of 89.45% and 89.98% respectively. The compactness of the discovered clusters by using AECRAM showed high silhouette width of 0.61% and 0.86% and 2.31% and 2.16% reduced value of Dunn index relative to K-means and EKM algorithms. It is clear from the data that AECRAM clustered dataset1 more efficiently than the other two algorithms used (Table 4).

Table 5: Cluster validation of the three clustering algorithms for dataset2 for 50 Runs

Algorithms	**Standard Deviation**	**F-measure (F1Score)**	**Silhouette Coefficient**	**Dunn Index**
K-Means	451.13	71.59%	0.307736	0.22221
Evolutionary K-Mean	356.98	71.98%	0.311123	0.22132
AECRAM	211.5	72.11%	0.319076	0.21172

Table 6: Cluster validation of the three clustering algorithms for dataset3 for 50 Runs

Algorithms	**Standard Deviation**	**F-measure (F1Score)**	**Silhouette Coefficient**	**Dunn Index**
K-Means	15.13	42.99%	0.1781457	0.02764293
Evolutionary K-Mean	14.96	50.08%	0.180100	0.02770122
AECRAM	11.65	59.19%	0.20	0.02743278

Moreover, for the dataset2 also, AECRAM identified clusters more accurately with reduced standard deviation of 53.11% and 40.75% as compared to K-means and EKM. Thus, identified clusters were accurate in terms of F-measure with the observed values of their F1Score to be 72.11% as compared to algorithms K-means and EKM that resulted in 0.72% and 0.18% less F1Score with regard to their respective clusters. The compactness of the discovered clusters by using AECRAM was observed to be with high silhouette width of 3.55% and 2.49% and 4.95% and 4.53% reduced values of Dunn index relative to the algorithms K-means and EKM. It can be stated again that clustering using AECRAM is more efficient with the dataset2 too than the clustering using the other two algorithms (Table 5).

In addition, for the dataset3 as well, AECRAM out-competed K-means and EKM algorithms in the identification of clusters more accurately as exemplified by much lower standard deviation values of 11.65 relative to 15.13 and 14.96 values observed with K-means and EKM algorithms respectively. The clusters identification were also precise in terms of their F-measure with F1Score of 59.19% whereas clusters resulting from K-means and EKM showed reduced F1Score of 27.36% and 15.39% respectively. The compactness of the discovered clusters using AECRAM was observed to have high silhouette width of 10.92% and 9.95% and 0.76% and 0.97% reduced value of Dunn index (Table 6). It is clear from this data too that the algorithm AECRAM more efficiently clustered dataset3 than the other two algorithms. In addition, the cluster validation process using AECRAM algorithm has the ability to discover more accurate clusters from multivariate datasets than other non-evolutionary and evolutionary algorithms. This could be due to the adaptive process implemented by this algorithm wherein each item in the dataset is migrated until it gets adapted in a cluster and the algorithm adapts for its inter-cluster distance as the fitment factor for a cluster under investigation.

IV. CONCLUSION

Cluster validation is an important step in the process of clustering. The clusters identified by a clustering algorithm can be validated by three (internal, external and relative) cluster validation methods. In the present work adaptive evolutionary algorithm (AECRAM) was observed to be superior over K-means and EKM algorithms in terms of precision and accuracy in cluster validation of multivariate datasets without missing values. AECRAM has ability to evolve from one cluster to a number of other precise clusters even if the prior knowledge of actual number of clusters in the dataset is not there. Intriguingly, for the dataset4, AECRAM also clustered the items that remain un-clustered or imperfectly clustered by the other algorithms used. A swot of the adaptive evolutionary clustering algorithm for the datasets with missing values and overlapped clusters is the subject of future study.

V. REFERENCE

[1] Cheng, D., Zhu, Q., Huang, J., Wu, Q., & Yang, L. (2018). A Novel Cluster Validity Index Based on Local Cores. IEEE transactions on neural networks and learning systems, Vol. (99), pp. 1-15.

[2] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package ‘nbclust’. Journal of statistical software, Vol. 61, pp. 1-36.

[3] Dimitriadou, E., Dolničar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. Psychometrika, Vol. 67(1), 137-159.

[4] Dua, D., & Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. School of Information and Computer Science.

[5] Ferraro, M. B., & Giordani, P. (2015). A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets and Systems, Vol. 279(1), pp. 1-16.

[6] Hämäläinen, J., Jauhiainen, S., & Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. Algorithms, Vol. 10(3), pp. 105.

[7] Handl J., Knowles J., and Kell D. (2005) Computational cluster validation in post-genomic data analysis, Bioinformatics, Vol. 21, pp. 3201–3212.

[8] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of intelligent information systems, Vol. 17(2-3), 107-145.

[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, Vol. 11(1), pp. 10-18.

[10] Jyoti, L., Ajay, K., Anupama, C., & Dharmesh, H. (2016, August). Auto-Evolving Clusters based on Rejection and Migration. In Proceedings of the International Conference on Advances in Information Communication Technology & Computing (p. 98). ACM.

[11] Lewis, D. D. (1998, April). Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg.

[12] Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on pattern analysis and machine intelligence, Vol. 24(12), 1650-1654.

[13] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization, Vol. 752(1), pp. 41-48.

[14] Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, Vol. 50(2), 159-179.

[15] Moshtaghi, M., Bezdek, J. C., Erfani, S. M., Leckie, C., & Bailey, J. (2018). Online cluster validity indices for streaming data. arXiv preprint arXiv:1801.02937.

[16] Pal, N. R., & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems, Vol. 3(3), 370-379.

[17] Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. International Journal of computers and communications, Vol. 5(1), pp. 27-34.

[18] Team, R. C. (2013). R: A language and environment for statistical computing.

[19] Theodoridis, S. and Koutroumbas, K. (2008) Pattern Recognition IEEE Transactions on Neural Networks, Vol. 19(2), pp. 376-376.

[20] Wang K., Wang B., and Peng L. (2009) CVAP: Validation for cluster analyses, Data Science Journal, Vol. 8(1), pp. 88–93.

[21] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[22] Wu, J., Chen, J., Xiong, H., & Xie, M. (2009). External validation measures for K-means clustering: A data distribution perspective. Expert Systems with Applications, Vol. 36(3), 6050-6061.

[23] Wu J., Xiong H., and Chen J. (2009) Adapting the right measures for k-means clustering, In 15th ACM SIGKDD Conference on Knowledge Discovery and DataMining (KDD'09), pp. 877-886

[24] Zhang, Y., Wang, W., Zhang, X., & Li, Y. (2008). A cluster validity index for fuzzy clustering. Information Sciences, Vol. 178(4), 1205-1218.

[25] Zhu, E., Wang, X., & Liu, F. (2019, February). A new cluster validity index for overlapping datasets. In Journal of Physics: Conference Series (Vol. 1168, No. 3, p. 032070). IOP Publishing.

[26] Breast Cancer Wisconsin Dataset. Available at: UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/support/Breast+Cancer+Wisconsin+(Diagnostic) (accessed 25th December 2018)

[27] Clustering Validation Statistics [www document] http://www.sthda.com/english/wiki/print.php?id=241 (accessed on 25th December 2018)

AUTHORS' PROFILE

Ms. Jyoti Lakhani is Assistant Professor in the Department of Computer Science, Maharaja Ganga Singh University, Bikaner. She has ten years of teaching experience. She is pursuing her Ph.D. research in Computer Science and addressing the colossal issue of design and implementation of evolutionary clustering algorithms and its application to biological sequence analysis. Her core research interests is centered on improving the understanding, design, and performance of evolutionary computing techniques, mainly through the application of data mining and other soft computing tools. She has published several national and international research papers in peer reviewed journals, proceedings and edited books. She has authored two text books. She is member of IEEE and ACM societies.

Professor Ajay Khunteta has 18 years of teaching and research experience. He completed his PhD on Distributed Systems. He has guided 34 M. Tech and 01 Ph. D students in the field of Distributed Algorithms, Information Security, Cloud Computing, Big Data and Image Processing. He has published more than 48 research papers in national and international Journals and conferences. He is Member of IEEE, ISTE and IEI.

Dr. Anupama Choudhary is working as a principal in Keens College, Bikaner, and Rajasthan, India. She has 22 years of teaching and research experience. She has completed her PhD on Database System models. She has published several research papers in national and international journals and conference proceedings. Her research interests involve database systems, artificial intelligence and network security.

Dr. Dharmesh Harwani began his academic career in 2002 as a JRF in the field of biological diversity of *Bradyrhizobium japonicum* bacteria and AMF for improved BNF in South Asia in an international research project (INCO-DEV, European Commission) involving Germany, Spain, Nepal and India. He is a Ph.D. and has associations with research team of MRDG at Indian Institute of Science, Bangalore as a DBT-Post Doctoral Fellow and is serving as an Assistant Professor at Maharaja Ganga Singh University, Bikaner. He has received several awards including DST and ICMR travel awards for academic visits to USA and Europe. He is currently member of the ASM-USA and AMI-India etc. His research activities are sponsored by national grants from DBT and DST, India. He has published more than 40 national and international papers in peer reviewed journals, proceedings, bulletins and book chapters.