VINAYAKA : A SEMI-SUPERVISED PROJECTED CLUSTERING METHOD USING DIFFERENTIAL EVOLUTION

Satish Gajawada[1] and Durga Toshniwal[2]

Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, Roorkee, India
[1]gajawadasatish@gmail.com , [2]durgafec@iitr.ernet.in

ABSTRACT

Differential Evolution (DE) is an algorithm for evolutionary optimization. Clustering problems have been solved by using DE based clustering methods but these methods may fail to find clusters hidden in subspaces of high dimensional datasets. Subspace and projected clustering methods have been proposed in literature to find subspace clusters that are present in subspaces of dataset. In this paper we propose VINAYAKA, a semi-supervised projected clustering method based on DE. In this method DE optimizes a hybrid cluster validation index. Subspace Clustering Quality Estimate index (SCQE index) is used for internal cluster validation and Gini index gain is used for external cluster validation in the proposed hybrid cluster validation index. Proposed method is applied on Wisconsin breast cancer dataset.

KEYWORDS

Semi-supervised Clustering, Projected Clustering, Differential Evolution, Hybrid Cluster Validation Methods

1. INTRODUCTION

Differential Evolution (DE) was proposed by Price and Storn [1] which is a evolutionary based optimization technique and based on a differential operator. Engineering problems like aerodynamic design, mechanical design optimization, design of digital filters and multiprocessor synthesis have been solved by DE [2]. Differential Evolution was applied for solving clustering problems [3]. But due to problems associated with high dimensional datasets DE based clustering methods proposed in literature may fail to find clusters in high dimensional datasets. The dataset may contain irrelevant dimensions. As the number of dimensions increases in dataset, the distance measures become increasingly meaningless. Each cluster may exist in different subspaces of dataset [4]. Subspace clusters which exist in subspaces of dataset can be found by subspace and projected clustering methods [5]. These methods may be applied for finding subspace clusters in different applications like metabolic screening, gene expression analysis, text documents and customer recommendation systems [6]. Hence there is need for DE based high dimensional data clustering methods which can find subspace clusters in high dimensional datasets.

The clustering results are evaluated using cluster validity indices. Cluster validation indices which use information present in the data are called as internal cluster validation indices. External cluster validation indices use external information that is available about the data [7]. The optimal clustering solution can be identified by executing a clustering algorithm several times with different input parameters each time and validating clusters obtained with a cluster validation

DOI : 10.5121/ijsea.2012.3406

index. The optimal clustering solution is the one which has the best value for cluster validation index [8]. Various cluster validation indices were defined in literature [9-12]. There exists no best cluster validation index which always gives better result compared to other indices. Better results can be obtained by fusion of various cluster validation indices compared to using single cluster validation index for getting optimal clustering solution. Internal validation indices like Davies-Bouldin index and Dunn index can be fused to validate clustering solutions for obtaining optimal clustering solution [13]. Using fusion of internal validation indices can give better results but available external information about the dataset is not used in the validation of clustering solution. Hybrid cluster validation indices are based on internal and external cluster validation indices. These indices use the available external information in the validation process in addition to intrinsic information present in the data [15].

Impurity of certain split in decision trees was measured in literature by using impurity measures like gini index, entropy index, classification error index and information gain ratio index. Gini index, Entropy index and Classification error index can be used in the gain criterion defined in [14] to get Gini gain index, Information gain index and Classification error gain index respectively. The quality of clustering solutions can be evaluated using these impurity measures [15].

Many cluster validation techniques are designed for evaluating clustering quality of full dimensional clustering methods. These techniques find difficulty to evaluate the clustering quality of subspace and projected clustering methods which find clusters present in subspaces of datasets. Subspace Clustering Quality Estimate index (SCQE index) was proposed in literature to evaluate the clustering quality of clustering methods which finds subspace clusters. SCQE index was defined based on Davies-Bouldin index [16].

In this paper we propose a hybrid cluster validation index for high dimensional datasets using SCQE index for internal cluster validation and Gini index gain for external cluster validation. We also propose a semi-supervised clustering method where DE optimizes the proposed hybrid cluster validation index.

The remainder of the paper is organized as follows: Related work is given in Section 2. Section 3 explains proposed work. Section 4 includes results and discussion. Conclusions are made in Section 5.

2. Related Work

Differential Evolution (DE) was applied for clustering by Das et al. [3]. In [17] document clustering was performed by using DE. Sudhakar et al. [18] clustered image datasets with DE. A comparision of data clustering by Particle swarm optimization (PSO) and DE techniques was made by Sai Hanuman et al. [19]. Simulation results showed that DE could provide better performance compared to PSO.

Due to various problems associated with clustering of high dimensional datasets the above given methods tend to fail for these datasets. Hence subspace and projected clustering methods were proposed to overcome the problems associated with clustering of high dimensional datasets. Approaches which start from full dimensional space to find relevant attributes of cluster are known as top-down approaches [6]. PROCLUS [20] randomly selects set of points from input dataset. A set of scattered medoids are obtained from the randomly selected points. Subspace is determined for each medoid and points are assigned to the medoids based on subspace identified. Clustering quality is increased by replacing the bad medoids with new medoids. This replacement of medoids is performed as long as quality of clustering increases. Clusters identified are used for finding relevant dimensions again. Reassignment of points to medoids is done based on relevant dimensions found. DOC [21] selects tentative cluster members and an arbitrary point for the cluster. Relevant attributes for cluster are identified by using projections of tentative cluster members and projection of selected arbitrary point on attributes. The procedure for finding single cluster is repeated several times and cluster with highest quality is taken. The other projected clusters are found in similar way. A specialized distance measure is used in PreDeCon [22]. Each point is associated with separate weights for all attributes. Relevant attributes are found by using variance of the points in a full-dimensional ε-neighborhood. Relevant attributes are given weight $k>>1$ and the other attributes are given weight 1. Clusters are then identified by using a full dimensional density based clustering method.

Cluster validation techniques were used for predicting the number of clusters in the dataset. The number of clusters in cancer tumor datasets is estimated by Bolshakova et al. [23]. The prediction of number of clusters was improved by a weighted voting technique. The number of clusters in artificial datasets was examined by Dimitriadou et al. [24] using 14 cluster validation indices. A new method to estimate number of clusters in the dataset was developed by Dudoit et al. [25]. A review of cluster validity measures available in literature was presented by Halkidi et al. [26]. A comparision of various internal and external validation indices was made by Erendira Rendon et al. [7]. Satish et al. [27] used Genetic algorithm to find optimal level of cutting the dendrogram obtained by hierarchical clustering on input dataset. A document clustering method based on cluster validation was presented by Zheng-Yu Niu et al. [28]. Multiple internal cluster validation indices was used by Krzysztof Kryszczuk et al. [13] to estimate the number of clusters. Significant gains in accuracy in estimating the number of clusters was obtained by fusion of multiple cluster validation indices. A weighted rank aggregation of various cluster validation measures was performed by Pihur et al. [29] using a Monte Carlo approach. Demiriz et al. [30] proposed a hybrid index based on gini index and Davies-Bouldin (DB) index. An effective framework based on clustering and classification was proposed by Patil et al [31]. Labels have been assigned through clustering. The noise points were eliminated by matching assigned labels with given labels in the pre-process step. Promising classification accuracy was obtained by proposed framework as compared to other methods found in literature. Classification accuracy of classifier can be improved by using internal information present in the data [31]. Hence the results of clustering methods which use external information for validation purposes can be improved by using the validation methods which are based on internal information about the data. A cluster validation metric known as Subspace Clustering Quality Estimate metric (SCQE metric) for evaluating subspace clusters was proposed by Urszula Markowska-Kaczmar et al. [16] as it is difficult to use cluster validation techniques designed for full dimensional clusters to evaluate subspace clusters.

3. PROPOSED WORK

In this section proposed VINAYAKA method is explained. Figure 1 shows the VINAYAKA method. Equation (1) shows proposed hybrid cluster validation index for evaluating subspace clusters obtained from proposed projected clustering method. W1 represents the weight given to internal cluster validation component and W2 represents the weight given to external cluster validation component. SCQE was proposed in [16] for evaluating quality of subspace clusters obtained from subspace clustering methods. Gain obtained from Gini index can be used for finding impurity of certain split in decision trees [14]. This gain criterion can be used for external cluster validation [15]. In the proposed hybrid SCQE-Gini Gain index, the evaluation of clustering solution is based on SCQE index and Gini Gain index. Similarly, we can obtain other hybrid cluster validation methods for evaluating quality of subspace clusters present in subspaces of high dimensional datasets.

$$\text{SCQE-Gini Gain index} = W1*\text{SCQE index}+ W2*\text{Gini Gain index} \qquad (1)$$

From Figure 1 we can observe that DE finds optimal centers of subspace clusters by optimizing a hybrid cluster validation index. In the fitness function, clusters are obtained by identifying relevant attributes based on neighbourhood of cluster centers and assignment of points to centers using relevant attributes found. Clusters obtained are used for finding relevant attributes again and re-assignment of points is done based on relevant attributes found to obtain subspace clusters. The subspace clusters obtained are validated using proposed hybrid cluster validation index. The value of hybrid cluster validation index is taken as fitness value of the vector. The algorithm returns optimal cluster centers of subspace clusters after termination.

```
Generate initial population

Evaluate below six steps  for each vector to obtain fitness values
        Find neighbourhood of cluster centers
        Identify relevant attributes of clusters
        Assign points to cluster centers using relevant attributes found in
        above step
        Use clusters identified in above step to find relevant attributes of
        clusters
        Reassign points to cluster centers using relevant attributes identified
        in above step
        Hybrid subspace cluster validation:
        a)  Find SCQE index
        b)  Find Gini gain index using class labels
        c)  Fitness= W1*(SCQE index) + W2* (Gini gain index)

LOOP until termination condition reached

        Get mutated vector
        Get trial vector

    Evaluate below six steps for each trial vector to obtain fitness values
        Find neighbourhood of cluster centers
        Identify relevant attributes of clusters
        Assign points to cluster centers using relevant attributes found in
        above step
        Use clusters identified in above step to find relevant attributes of
        clusters
        Reassign points to cluster centers using relevant attributes identified
        in above step
        Hybrid subspace cluster validation:
        a)  Find SCQE index
        b)  Find Gini gain index using class labels
        c)  Fitness= W1*(SCQE index) + W2* (Gini gain index)

Replace target vector with trial vector if trial vector has better fitness value

END LOOP
RETURN optimal subspace cluster centers
```

Figure 1. The proposed VINAYAKA method

4. EXPERIMENTAL RESULTS

In this section we present the results obtained for some synthetic datasets with a hybrid cluster validation index proposed in [15]. We also present the results obtained for Wisconsin breast cancer data [32] with proposed method.

k-means clustering has been used to obtain clustering solutions from synthetic datasets. The clustering solutions obtained for various values of number of clusters parameter are validated with a hybrid cluster validation index and an internal validation index. First dataset is organized into 15 clusters in 3 dimensional space. Each cluster is given a separate class label. Hence there are 15 classes in the first dataset. There are 5 groups of 3 close clusters in the dataset. Second dataset is created by giving a single class label to each group of 3 close clusters. So, second dataset has 5 labels because there are 5 groups each of which is given a separate class label. Class labels have been assigned to only a part of the datasets because complete external information may not always be available. Both datasets contain labelled data part and unlabelled data part. Internal index has been calculated using all points in the dataset. External index has been calculated using labelled data part of the dataset. Results have been obtained using Dunn index and hybrid Dunn-CEGR index [15] with equal weights on two datasets. Results obtained are discussed below.

Figure 2 shows the index plot of Dunn index for both the datasets. As both datasets differ only in class labels Dunn index plot will be same for both the datasets. Figure 3 shows the index plot of dataset 1 with hybrid Dunn-CEGR index. Figure 4 shows the index plot of dataset 2 with hybrid Dunn-CEGR index.

Table 1 shows optimal number of clusters obtained with Dunn index and hybrid index when applied on two datasets. Table 1 is created by taking minimum (optimal values) of cluster validation index values. These optimal index values can be obtained from figures Figure 2 to Figure 4.

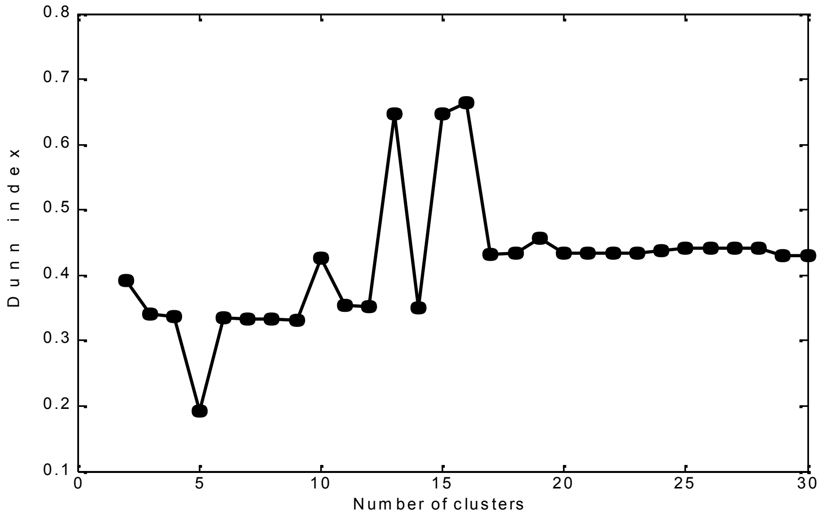

Figure 2. Index plot of dataset 1 and dataset 2 with Dunn index

Table 1. Optimal number of clusters obtained

Dataset / index	Dunn index	Hybrid Dunn-CEGR index
Dataset 1	5	14
Dataset 2	5	5

From Table 1 we can observe that Dunn index gave 5 clusters as optimal for both datasets. This is because there are 5 groups of 3 close clusters and we get 5 clusters by clubbing 3 close clusters to a single cluster and Dunn index value is optimal at 5 clusters. Result obtained by Dunn index is correct for dataset 2. For dataset 1 difference between expected result (15 clusters) and obtained result (5 clusters) is 10 clusters. Hybrid Dunn-CEGR index gave 14 clusters for dataset 1 and hence error in number of clusters is just 1 cluster. For dataset 2 hybrid index gave correct result of 5 clusters. This is because we are using some class labels available for cluster validation in addition to Dunn index.

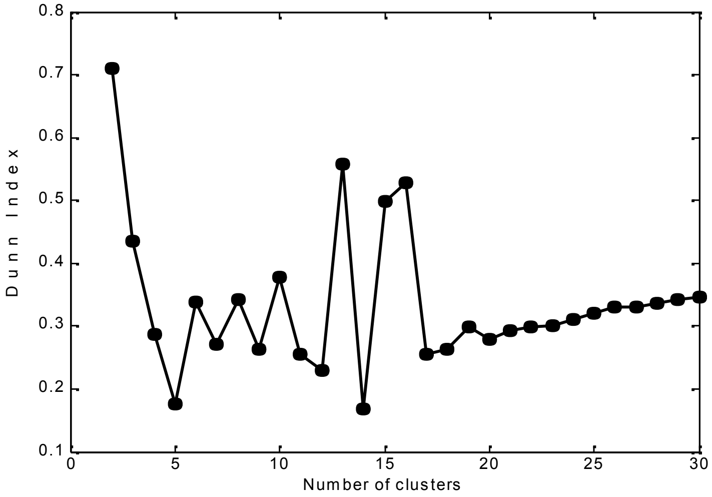

Figure 3. Index plot of dataset 1 with hybrid Dunn-CEGR index

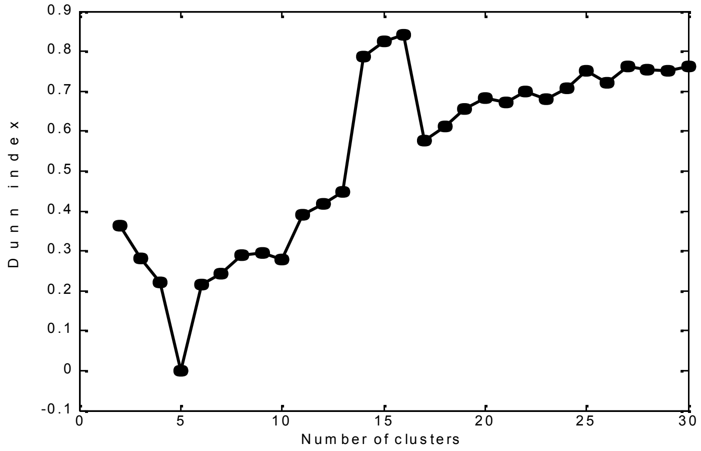

Figure 4. Index plot of dataset 2 with hybrid Dunn-CEGR index

Both datasets used differ only in class labels. When there are group of close clusters in the dataset then all these close clusters may be clubbed and considered as single large cluster or each cluster in the group of close clusters can be considered as separate cluster. But using only internal information present in the dataset will give same result for both cases. Hence using little external information available for cluster validation in addition to internal information present in data can give different result for both cases.

Table 2. Matching points between output and input clusters of Wisconsin breast cancer data for average subspace dimensions 6

Cluster	A	B
1	435	22
2	9	217

Table 3. Relevant dimensions of output clusters of Wisconsin breast cancer data

Output	Dimensions
1	2,3,4,5,6,8,9
2	2,3,5,6,9

The clusters obtained by using VINAYAKA method are referred as output clusters and clusters present in the dataset are referred as input clusters. Table 2 shows matching points between output clusters {1, 2} and input clusters {A, B} of Wisconsin breast cancer data for 6 average subspace dimensions. Input clusters A and B correspond to two classes present in Wisconsin breast cancer data. Each row has a large value compared to other value which shows that each output cluster matched to input cluster present in the data. Output cluster 1 matched to input cluster A and output cluster 2 matched to input cluster B. Output cluster 1 contains 22 misclassified points and output cluster 2 contains 9 misclassified points.

Table 3 gives the relevant dimensions of output clusters identified by using proposed method for Wisconsin breast cancer data for average number of subspace dimensions 6. First output cluster has relevant dimensions {2,3,4,5,6,8,9} and second output cluster has relevant dimensions {2,3,5,6,9}.

5. CONCLUSIONS

In this paper we proposed VINAYAKA, a semi-supervised projected clustering method using Differential Evolution optimization technique. We also proposed a hybrid cluster validation index for evaluating quality of subspace clusters obtained from projected clustering methods. In the VINAYAKA projected clustering method, DE obtains optimal cluster centers of subspace clusters by optimizing a hybrid cluster validation index. The hybrid cluster validation index proposed in this paper is based on SCQE index and Gini Gain index. The proposed semi-supervised projected clustering method is applied on Wisconsin breast cancer dataset to find subspace clusters present in this dataset. It has been observed that 95.46 percentage of points in the dataset are correctly classified based on subspace clusters identified by using VINAYAKA method. Our future work includes creation of several new classification methods based on semi-supervised projected clustering method proposed in this work.

ACKNOWLEDGEMENTS

The corresponding author of this paper would like to thank everyone who encouraged him during the research work.

REFERENCES

[1] R. Storn, and K. Price, “Differential Evolution - a Simple and Efficient Heuristic for Global Optimization over Continuous Spaces,” Journal of Global Optimization, Kluwer Academic Publishers, Vol. 11, 1997, pp. 341 - 359.

[2] Zhihua Cai, Wenyin Gong, Charles X. Ling, and Harry Zhang, “A clustering-based differential evolution for global optimization,” Applied Soft Computing, Volume 11, Issue 1, January 2011, pp. 1363-1379.

[3] S. Das, A. Abraham, and A. Konar, “Automatic Clustering Using an Improved Differential Evolution Algorithm,” IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 38, No. 1, 2008, pp. 218-237.

[4] L. Parsons, E. Haque, and H. Liu, “Subspace clustering for high dimensional data: A review,” SIGKDD Explor, Vol. 6, 2004, pp. 90-105.

[5] G. Moise, A. Zimek, P. Kroger, H.P. Kriegel, and J. Sander, “Subspace and projected clustering: experimental evaluation and analysis,” Knowl. Inf. Syst., Vol. 3, 2009, pp. 299-326.

[6] H.P. Kriegel, P. Kroger, and A. Zimek, “Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering,” ACM Trans. Knowl. Discov. Data., Vol 3, 2009, pp. 1-58.

[7] ErendiraRendon, Itzel Abundez, Alejandra Arizmendi, Elvia M. Quiroz, “Internal versus External cluster validation indexes,” International Journal of Computers and Communications, Vol. 5, No. 1, 2011, pp. 27-34.

[8] Bolshakova, N., Azuaje, F., “Machaon CVE: cluster validation for gene expression data,” Bioinformatics, Vol. 19, No. 18, 2003, pp. 2494-2495.

[9] Rousseeuw, P. J., “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” J. Comp App. Math, Vol. 20, 1987, pp. 53-65.

[10] Dunn, J., “Well separated clusters and optimal fuzzy partitions,” J. Cybernetics, Vol. 4, 1974, pp. 95-104.

[11] Davies, D.L., Bouldin, D.W., “A cluster separation measure,” IEEE Transactions on Pattern Recognition and Machine Intelligence, Vol. 1, No. 2, 1979, pp. 224-227.

[12] Hubert, L., Schultz, J., “Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie,” Vol. 29, 1976, pp. 190-241.

[13] Krzysztof Kryszczuk, Paul Hurley, “Estimation of the number of clusters using multiple clustering validity indices,” MCS 2010, LNCS 5997, 2010, pp. 114-123.

[14] Pang Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining. Pearson Education, 2009.

[15] Satish Gajawada and Durga Toshniwal, “Hybrid Cluster Validation Techniques,” Proceedings of the Second International Conference on Computer Science, Engineering & Applications (ICCSEA 2012), 2012, pp 267-273.

[16] Urszula Markowska-Kaczmar and Arletta Hurej, “Evaluation of Subspace Clustering Quality,” HAIS 2008, LNAI 5271, 2008, pp. 400–407.

[17] A. Abraham, S. Das, and A. Konar, “Document Clustering Using Differential Evolution,” IEEE Congress on Evolutionary Computation, 2006, pp. 1784 – 1791.

[18] G. Sudhakar, Polinati Vinod Babu, Suresh Chandra Satapathy, and Gunanidhi Pradhan, “Effective Image Clustering with Differential Evolution Technique,” Int. J. of Computer and Communication Technology, Vol. 2, No. 1, 2010, pp. 11-19.

[19] A. Sai Hanuman, Dr. A. Vinaya Babu, Dr. A. Govardhan, and Dr. S. C. Satapathy, “Data Clustering using almost parameter free Differential Evolution technique,” International Journal of Computer Applications, Vol. 8, No. 13, October 2010, pp. 1–7.

[20] C. C. Aggarwal, C. M. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park, “Fast algorithms for projected clustering,” In Proceedings of the ACM International Conference on Management of Data (SIGMOD), 1999, pp. 61-72.

[21] C. M. Procopiuc, M. Jones, P. K. Agarwal, and T.M. Murali, “A Monte Carlo algorithm for fast projective clustering,” Proceedings of the ACM International Conference on Management of Data (SIGMOD), 2002, pp. 418-427.

[22] C. Bohm, K. Kailing, H. P. Kriegel, and P. Kroger, “Density connected clustering with local subspace preferences,” Proceedings of the 4th International Conference on Data Mining (ICDM), 2004, pp. 27-34.

[23] Bolshakova, N., Azuaje, F., “Estimating the number of clusters in DNA microarray data,” Methods of Information in Medicine, 2006.

[24] Dimitriadou, E., Dolnicar, S., Weingessel, A., “An examination of indexes for determining the Number of Cluster in binary data sets,” Psychometrika, Vol. 67, No. 1, 2002, pp. 137-160.

[25] Dudoit, S., Fridlyand, J., “A prediction-based resampling method for estimating the number of clusters in a dataset,” Genome Biology,Vol. 3, No. 7, 2002.

[26] Halkidi, M., Batistakis, Y., Vazirgiannis, M., “On Clustering Validation Techniques. Intelligent Information Systems Journal,” Vol. 17, No. 2, 2001, pp. 107-145.

[27] Satish, G., Durga, T., Nagamma, P., Kumkum, G., “Optimal Clustering Method Based on Genetic Algorithm,” International conference on soft computing for problem solving, Advances in Intelligent and Soft Computing, Volume 131, 2012, pp 295-303.

[28] Zheng-Yu Niu, Dong-Hong Ji, Chew-Lim Tan, “Document Clustering Based on Cluster Validation,” Proceedings of the Thirteenth ACM conference on Information and knowledge management CIKM 04, 2004.

[29] Pihur, V., Datta, S.,Datta, S., “Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach,” Bioinformatics, Vol. 23, No. 13, 2007, pp. 1607-1615.

[30] Demiriz, A., Bennett, K.P., Embrechts, M.J., “Semi-supervised clustering using genetic algorithms,” Artificial neural networks in engineering, 1999, pp. 1-20.

[31] Patil, B.M., Joshi, R. C., Durga, T., “Effective framework for prediction of disease outcome using medical datasets: clustering and classification,” Int. J. Computational Intelligence Studies, Vol. 1, No. 3, 2010.

[32] Frank, A. and Asuncion, A. UCI Machine Learning Repository, available at http://archive.ics.uci.edu/ml, Irvine, CA: University of California, School of Information and Computer Science, 2010.