International Journal of Soft Computing 9 (3): 131-137, 2014
ISSN: 1816-9503


# An Initialization Method for K-Means Algorithm Using Binary Search Technique

Yugal Kumar and G. Sahoo
Department of Information Technology, Birla Institute of Technology,
Mesra, Ranchi Jharkhand, India

**Abstract:** K-Means algorithm is most popular partition based algorithm that is widely used in data clustering. A lot of algorithms have been proposed for data clustering using K-Means algorithm due to its simplicity, efficiency and ease convergence. In spite this K-Means algorithm has some drawbacks like initial cluster centers. In this study, a new method is proposed to address the initial cluster centers problem in K-means based on binary search technique. The initial cluster centers is obtained using Binary Search Method and the newly generated cluster centers are used as initial cluster centers in K-means to gain optimal cluster centers in dataset. The performance of the Proposed algorithm is tested on the two benchmark dataset iris and wine that are downloaded from the UCI machine learning repository and compare the proposed method with Random, Hartigan and Wang, Ward, Build, Astrhan, Minkowaski ward and IWKM Method in which proposed method with K-means provides 82.93 and 68.94 accuracy rate and intra cluster distance is 105.72 and 18059.81 with iris and wine datasets as well as proposed method with IWKM provides 96.7 and 95.8 accuracy rate.

**Key words:** Clustering, cluster centers, K-means, binary search, accuracy rate

## INTRODUCTION

Data clustering is an important technique for data analysis which can be used to discover the similarity or dissimilarity between groups of items in a dataset such that items in one group are more similar than other groups and vice versa (Jain *et al*., 1999). Mathematically the clustering problem can be defined by number of attributes and number of partitions in the dataset. For a given dataset the M number of attributes can be defined as:

$$X = \{x_1, x_2, x_3, ..., x_m\} \quad \text{where, } x_i \in R^N$$

and the P number of partitions on the dataset can be defined as:

$$P = \{P_1, P_2, P_3, ..., P_k\} \quad \text{where,}$$

$$\forall i \neq j \;\; P_i \cap P_j = \varnothing, \;\; \bigcup_{i=1}^{k} P_i = X, \;\; \forall i \;\; P_i = \varnothing$$

So, the clustering problem can be viewed as searching problem that can searches a particular partition with minimum criterion function. Sum of squared error is the most common criterion function that can be used to search a particular partition in clustering task. Sum of squared error function can be defined as:

$$SSE(X, P) = \sum_{i=1}^{k} \sum_{x_j \in P_i} \|x_j - p_i\|^2$$

Large number of algorithms have been developed by various researchers for data clustering task. In the clustering domain K-means is the oldest and probably the most popular algorithm proposed is proposed by (MacQueen, 1967). It is easy to implement and it is fast and sensitive. However, the K-Means algorithm has some drawbacks (Selim and Ismail, 1984; Kao *et al*., 2008; Jain, 2010). These are:

- Lack of knowledge how to treat with inappropriate and clatter attributes
- Lack of universal method how to choose the initial location of cluster centroids
- No information about number of clusters in the dataset
- Stuck in local optima

To overcome the drawbacks of K-Means algorithm, a lot of research have been done by various researchers. To enhancement and feature weighting in K-Means algorithm, a lot of research has been done by Modha and Spangler (2000, 2003) and Dhillon and Modha (2001). To overcome the attribute selection problem in K-Means algorithm, Huang *et al*. (2008) have proposed an automated weighted method for attribute selection and called it weighted K-Means algorithm. The sum of within cluster dispersion is used to calculate the weight of attributes and the attributes with lower weight is removed from the set of attributes. Another issue related to the K-Means algorithm is how many numbers of clusters exist

**Corresponding Author:** Yugal Kumar, Department of Information Technology, Birla Institute of Technology, Mesra, Ranchi Jharkhand, India

in a dataset and initialization of initial cluster centers. In real life clustering problems it is quite difficult to choose the number of clusters present in final result (Sneath and Sokal, 1973; Everitt, 1979). A large numbers of procedures have been developed to determine the number of clusters present in the dataset (Dubes and Jain, 1979; Milligan, 1981; Perruchet, 1983). These procedures are divided in various categories such as variance based approach (Hartigan, 1975; Krzanowski and Lai, 1988; Sugar and James, 2003) structural approach (Hubert and Levin, 1976; Milligan and Cooper, 1985; Shen *et al*., 2005) consensus distribution approach (Monti *et al*., 2003; Mirkin, 2005; Kuncheva and Vetrov, 2006), hierarchical approach (Duda and Hart, 1973; Milligan and Cooper, 1985; Ishioka, 2005; Feng and Hamerly, 2007) and resampling approach (Mirkin, 2005; Dudoit and Fridlyand, 2002; Minaei-Bidgoli *et al*., 2004; Mufti *et al*., 2005).

In this study, a new method is proposed to enhance the initialization problem of K-Means algorithm because the convergence result of K-Means algorithm is highly dependent on the initial cluster centers. If the initial cluster centers are not chosen properly then the local optimum problem will be exist in K-means. The good convergence result is directly proportional to the good cluster centers. Hence, the proposed method addresses the initialization as well as local optimum issues of K-means.

**Literature review:** De Amorim and Mirkin (2012) have developed Minkowaski Weighted K-Means (MW K-Means) algorithm and intelligent Minkowaski Weighted K-means (iMW K-means) algorithm to address the attribute selection and initial location of cluster center problems in which minkowski metric is used as distance measure and initial cluster center is specified using anomalous clusters. De Amorim and Komisarczuk (2012) have used six different Centroid Initialization Methods with Minkowski Weighted K-Means (MWK-Means) algorithm to evaluate which method gives better performance. These methods are compared on the behalf of the accuracy and processing time in which Ward Method provides good results. To reduce the dependency of the K-means on initial centroid (Chan *et al*., 2006) have applied the greedy elimination method with K-means to generate consistent and optimal clusters center in gene expression data and this method produces better results as compare to standard K-means and fast greedy Incremental Method. To initialize the k clusters points in K-Means algorithm, Bradley and Fayyad (1998) have used minimizing concave function as bilinear program to evaluate the k number of clusters centers. A heuristic function based on the mode of the joint probability density function has applied by Bradley *et al*. (1997) with K-Means algorithm to generate good initial clusters points. Cao *et al*. (2009) have proposed a new method based on the cohesion and coupling degree between adjacent items using rough set model to determine the initial clusters center using K-means. To overcome the dependency of K-Means algorithm on the initial clusters centroids (instead of random generation), Likas *et al*. (2003) have developed a global K-Means algorithm based on the deterministic global optimization and K-means in which K-means implemented as local search algorithms. To generate good initial cluster centers, Lu *et al*. (2008) have applied the hierarchical clustering approach with K-Means algorithm and this method required less iteration time and higher convergence speed but the method has some drawback such that the values of attributes must be numeric if the values are non numeric then these values must be converted into numeric values. According to Meila and Heckerman (1998) and De Amorim and Komisarczuk (2012), there are large number of methods exist to refine and initialization of the clusters centers in K-means but at present there is not a single method that can be recognized as universal method to generate initial cluster centers. Table 1 provides the list of well known

Table 1: Different initialization method to generate initial cluster center

| Initialization method | Cluster center |
|---|---|
| Random | A random function is used to generate the cluster center, i.e., rand () |
| Hartigan | Generate the cluster center using the following equation: 1+(k-1)×[N/K], where, k = 1, 2, 3,..., K and N = numbers of instance and the numbers of instance must be sorted according to center of gravity |
| Ward | Generate the initial center based on the ward criteria: $\text{dist}(a,b)=\frac{n_a n_b}{n_a+n_b}\| x_a-x_b \|^2$ Where, $n_a$ = number of points in cluster a, $n_b$ = number of points in cluster b, $x_a$ = cluster points of cluster a, $x_b$ = cluster points of cluster b |
| Build algorithm | Generate the initial center based on the following equation: $M_y=\sum_{y\in(D-C)} d(j,c_1)-d(y,j)$ Where, C = $\{c_1\}$ is the close to median, D = set of all instances; $M_y$ = Maximum value of data instance for $c_1$ |
| Astrahan | Generate the initial cluster using the density of data. The density of data is calculated by given equation in the dataset: $d_i=\frac{1}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\| y_i-y_j \|$ |

different initialization methods that are used with K-Means algorithm as well as criteria to generate the initial cluster centers.

## MATERIALS AND METHODS

**Proposed initialization method:** In this study, an initialization method is proposed for K-Means algorithm. The proposed method is used to generate the initial cluster centers rather than random or user specified cluster centers. The proposed initialization method is based on the algorithm proposed by Hatamlou (2012) to obtain most favorable cluster points. In the proposed method, the initial cluster points are generated by using the unique property of Binary Search algorithm to find the value of middle item in a given list, i.e.:

$$A[mid] = \frac{A[beg] + A[end]}{2} \quad (1)$$

The above property of binary search is modified to generate the initial cluster point for K-Means algorithm:

- A[beg] is replaced by A[max]
- A[end] is replaced by A[min]
- 2 is replaced by K, numbers of clusters
- A[mid] is replaced by any variable such as M
- Plus symbol is replaced by minus symbol

Now a new equation is generated using Eq. 1:

$$M = \frac{A(max) - A(min)}{K} \quad (2)$$

The generalization of Eq. 2 can be written as:

$$M_i = \frac{\max(A_i) - \min(A_i)}{K} \quad (3)$$

Equation 3 is used to calculate the value of the variable M that specifies the range of initial cluster centers but not give the cluster centers. The cluster centers for K-means algorithm are generated using given Eq. 4:

$$C_k = \min(A_i) + (K-1)M \quad (4)$$

Consider an example dataset D that is given in Table 2. The given dataset is applied with proposed method to get the initial cluster points. This dataset is

Table 2: Example dataset D to generate the initial cluster center

| Objects | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 | x11 | x12 | x13 | x14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1.1 | 1.3 | 1.2 | 3.2 | 2.8 | 2.9 | 2.0 | 1.9 | 2.2 | 8.0 | 7.4 | 7.2 | 9 | 8.8 |
| B | 4.3 | 3.9 | 3.8 | 4.8 | 3.9 | 3.7 | 3.6 | 3.3 | 3.2 | 3.2 | 5.9 | 3.8 | 4 | 6.9 |

consist total number of instances (N) = {14}, No. of attributes (i) = {2} and No. of clusters (K) = {3}. The working of proposed method is given as: calculate the maximum and the minimum values of each attribute in the dataset:

$$\text{maximum} = (9, 6.9) \text{ and minimum} = (1.1, 3.2)$$

Calculate the value of M as:

$$M = \left\{\frac{(9-1.1)}{3}, \frac{(6.9-3.2)}{3}\right\}$$
$$= (2.66, 1.23)$$

Generate the initial cluster centers for initialization as:

$$C_1 = (1.1 + (1-1)\times 2.66, 3.2 + (1-1)\times 1.23)$$
$$= (1.1, 3.2)$$
$$C_2 = (1.1 + (2-1)\times 2.66, 3.2 + (2-1)\times 1.23)$$
$$= (3.76, 4.43)$$
$$C_3 = (1.1 + (3-1)\times 2.66, 3.2 + (3-1)\times 1.23)$$
$$= (6.32, 5.66)$$

The newly generated cluster centers are used as initial cluster centers for K-Means algorithm and address the initialization problem of K-means. The proposed algorithm can be defined as:

Input (Dataset (d) and k)
Step 1: Set the number of clusters (k) where k = 1, 2, 3… m
Step 2: Generate the range of the initial centroids using following:
$M = (\max(d_j) - \min(d_j))/k$ where j = 1, 2, 3,..., n
Step 3: Obtain the initial cluster centers using the following equation:

$$C_k = \min(A_i) + (K-1)^M$$

Step 4: Calculate the Euclidean distance as similarity measure of each attribute $x_j$ and assigned to cluster center $C_k$ using following equation:

$$D = \min(\| X_i - C_k \|^2)^{1/2}$$

Step 5: Recalculate the centers for each cluster centers using the equation in step 4 until the cluster centers are changed
Step 6: Quit and return the final cluster centers

Figure 1 shows the flow chart of proposed method with K-means.

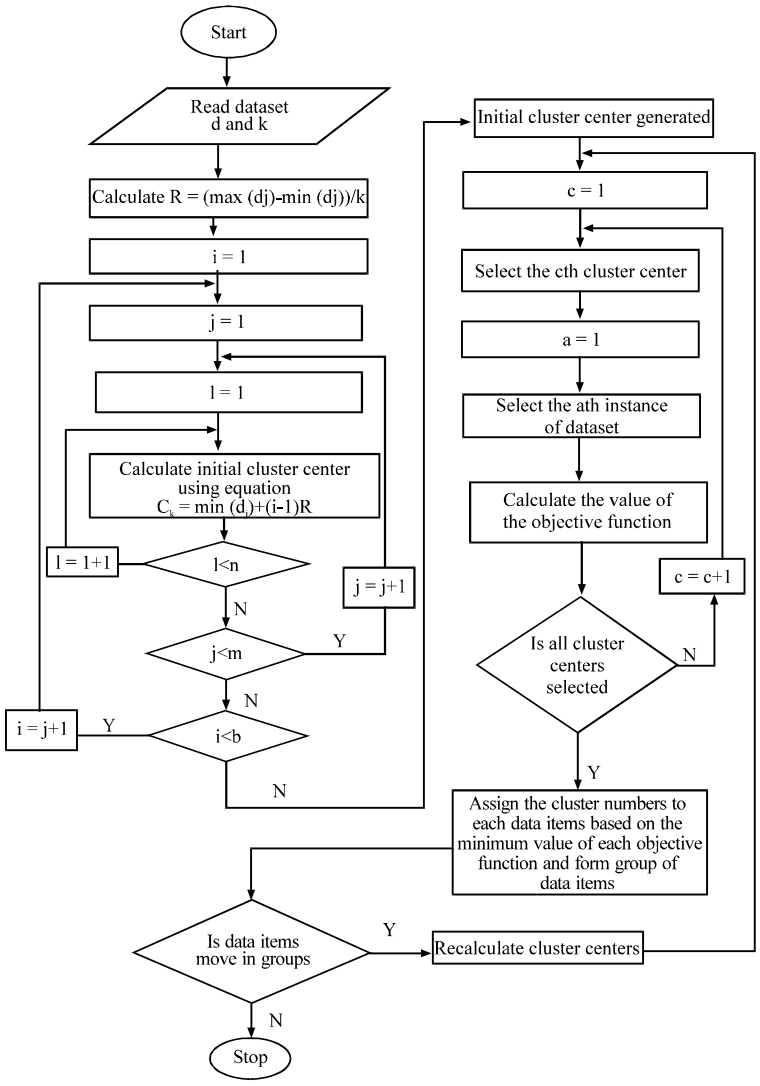


Fig. 1: Flow chart of the Proposed algorithm for initialization of K-means

## RESULTS AND DISCUSSION

The results are obtained on a system with Intel core i3 processor and windows 8 operating system. MATLAB 2010 (Math works) environment is used to code the proposed initialization method with simple K-means and Minkowaski Weighted K-Means (WKM) as well as Random, Hartigan, Ward, Build algorithm and Astrahan Initialization Methods with twenty runs of each algorithm. To ensure the efficiency of the proposed initialization method Iris and wine datasets are used and evaluate the performance of proposed method with Random, Hartigan, Ward, Build algorithm and Astrahan Initialization Methods using accuracy, intra cluster distance and processing time parameters. Accuracy has been defined here as the ratio of the total number of correctly predicted instances and the total number of instances. In clustering problem accuracy is defined as:

$$\text{Accuracy} = \sum_{i=1}^{k} \frac{a_i}{N}$$

Intra cluster distance is a quality parameter for clustering problem that can be defined as the sum of distances between instances within a cluster and the center points of cluster. Minimum sum of intra cluster distance indicates the good quality of cluster. Processing time is the amount of time required to execute the algorithm.

Table 3: Iris and wine datasets information

| Name | Total instance | Attributes | Class |
|---|---|---|---|
| Iris | 150 | 4 | 3 (50, 50, 50) |
| Wine | 178 | 13 | 3 (59, 71, 48) |

Table 4: Comparisons of different initialization methods and proposed initialization method for K-means algorithm

| | Iris | | Wine | |
|---|---|---|---|---|
| Method/parameters | Accuracy | Intra cluster distance | Accuracy | Intra cluster distance |
| Random | 82.60 | 106.50 | 68.88 | 18061.00 |
| Hartigan and Wang | 81.67 | 107.46 | 68.43 | 18061.46 |
| Minkowaski Ward | 82.43 | 106.26 | 68.27 | 18061.73 |
| Build | 81.92 | 106.89 | 68.56 | 18061.37 |
| Artshan | 82.46 | 106.43 | 68.63 | 18061.21 |
| Proposed method (BS) | 82.93 | 105.72 | 68.94 | 18059.81 |

Table 5: Comparisons of different initialization methods and proposed initialization method for IWKM algorithm

| | Iris | | | Wine | | |
|---|---|---|---|---|---|---|
| Method/parameters | Accuracy | P | Time (sec) | Accuracy | P | Processing time (sec) |
| Random | 96.7 | 1.2 | 0.40±2 | 96.1 | 2.3 | 2.40±1.0 |
| Hartigan and Wang | 96.7 | 1.1, 1.2 | 0.12 | 94.9 | 1.9, 2.2 | 4.74 |
| Minkowaski Ward | 96.7 | 1.1 | 0.43 | 95.5 | 1.9, 2.0 | 1.19 |
| Build | 96.7 | 1.1, 1.2 | 0.7 | 94.9 | 2.6 | 2 |
| Artshan | 96.7 | 1.2 | 0.99 | 95.5 | 2.5 | 3.29 |
| IWKM | 96.7 | 1.2 | 0.51 | 94.9 | 1.2 | 2.16 |
| Proposed method (BS) | 96.7 | 1.2 | 0.46 | 95.8 | 1.2, 1.4 | 1.46 |

**Dataset information:** Iris and wine datasets are used to measure the performance of the proposed initialization method. Table 3 provides the detailed information of iris and wine datasets.

Table 4 shows the comparison of proposed method and other five initialization methods with simple K-Means algorithm using accuracy and intra cluster distance parameters. From the Table 4, it is concluded that the proposed approach is obtained high accuracy rate (82.93, 68.94) and low intra cluster distance (105.72, 18059.81) for iris and wine dataset that shows the significance of proposed method to initialization of centroid for K-Means algorithm.

Table 5 provides the comparison of proposed method and six other initialization methods with Minkowaski WKM algorithm using accuracy, value of P and time parameters. From the Table 5, it is observed that proposed approach obtains high accuracy rate (95.8) with wine dataset using Minkowaski WKM algorithm while with iris dataset all initialization algorithms exhibits same behavior in terms of accuracy. On the analysis of time parameter, it is concluded that minowaski ward method require less processing time (1.19) among all methods for wine data set while Hartigan and Wang method require less processing time (0.12) for iris dataset. From the above study, it is concluded that proposed approach provides good results among all other initialization methods with simple K-means as well as Minkowaski WKM.

## CONCLUSION

This study focuses on the initialization of initial cluster centers problem in K-Means algorithm. To address the initialization problem, a binary search based initialization method with K-Means algorithm is proposed to initialize the initial cluster points for K-Means algorithm. In Proposed algorithm, initial cluster centers are obtained with the help of binary search based method and after that K-Means algorithm is applied. Performance of the Proposed algorithm is evaluated with some datasets that are downloaded from UCI repository and compared with five well known Cluster Centers Initialization Methods for K-Means algorithm. The performance of the Proposed algorithm is better than all other methods. The Binary Search Based Method is also used with Minkowaski WKM algorithm to generate the initial cluster points and the performance of this method is compared with six cluster centers initialization methods in which proposed method performs better.

## REFERENCES

Bradley, P.S. and U.M. Fayyad, 1998. Refining initial points for K-means clustering. Proceedings of the 15th International Conference on Machine Learning, July 24-27, 1998, Morgan Kaufmann, San Francisco, pp: 91-99.

Bradley, P.S., O.L. Mangasarian and W.N. Street, 1997. Clustering Via Concave Minimization. In: Advances in Neural Information Processing Systems, Mozer, M.C., M.I. Jordan and T. Petsche (Eds.). MIT Press, Cambridge, MA, USA., pp: 368-374.

Cao, F., J. Liang and G. Jiang, 2009. An initialization method for the K-Means algorithm using neighborhood model. Comput. Math. Appl., 58: 474-483.

Chan, Z.S.H., L. Collins and N. Kasabov, 2006. An efficient greedy K-Means algorithm for global gene trajectory clustering. Expert Syst. Appl., 30: 137-141.

De Amorim, R.C. and B. Mirkin, 2012. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. Pattern Recognit., 45: 1061-1075.

De Amorim, R.C. and P. Komisarczuk, 2012. On initializations for the minkowski weighted K-means. Proceedings of the 11th International Conference on Advances in Intelligent Data Analysis XI, October 25-27, 2012, Helsinki, Finland, pp: 45-55.

Dhillon, I.S. and D.S. Modha, 2001. Concept decompositions for large sparse text data using clustering. Mach. Learn., 42: 143-175.

Dubes, R. and A.K. Jain, 1979. Validity studies in clustering methodologies. Pattern Recognit., 11: 235-254.

Duda, R. and P. Hart, 1973. Pattern Classification and Scene Analysis. John Wiley and Sons, New York.

Dudoit, S. and J. Fridlyand, 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol., Vol. 3. 10.1186/gb-2002-3-7-research0036.

Everitt, B.S., 1979. Unresolved problems in cluster analysis. Biometrics, 35: 169-181.

Feng, Y. and G. Hamerly, 2007. PG-Means: Learning the Number of Clusters in Data. In: Advances in Neural Information Processing Systems, Scholkopf, B., J.C. Platt and T. Hofmann (Eds.). MIT Press, Cambridge, MA., USA., ISBN-13: 9780262195683, pp: 393-400.

Hartigan, J.A., 1975. Algorithm CHAID. Clustering Algorithms. John Wiley and Sons, New York.

Hatamlou, A., 2012. In search of optimal centroids on data clustering using a binary search algorithm. Pattern Recognit. Lett., 33: 1756-1760.

Huang, J.Z., J. Xu, M. Ng and Y. Ye, 2008. Weighting Method for Feature Selection in K-Means. In: Computational Methods of Feature Selection, Liu, H. and H. Motoda (Eds.). Chapman and Hall, New York, pp: 193-209.

Hubert, L.J. and J.R. Levin, 1976. A general statistical framework for assessing categorical clustering in free recall. Psychol. Bull., 83: 1072-1080.

Ishioka, T., 2005. An expansion of X-means for automatically determining the optimal number of clusters. Proceedings of the International Conference on Computational Intelligence, July 4-6, 2005, Calgary, AB., Canada, pp: 91-96.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett., 31: 651-666.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surveys, 31: 264-323.

Kao, Y.T., E. Zahara and I.W. Kao, 2008. A hybridized approach to data clustering. Expert Syst. Appl., 34: 1754-1762.

Krzanowski, W.J. and Y.T. Lai, 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics, 44: 23-34.

Kuncheva, L.I. and D.P. Vetrov, 2006. Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Trans. Pattern Anal. Mach. Intell., 28: 1798-1808.

Likas, A., N. Vlassis and J.J. Verbeek, 2003. The global K-Means Clustering algorithm. Pattern Recognit., 36: 451-461.

Lu, J.F., J.B. Tang, Z.M. Tang and J.Y. Yang, 2008. Hierarchical initialization approach for K-Means clustering. Pattern Recogn. Lett., 29: 787-795.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proc. Berkeley Symp. Math. Statist. Prob., 1: 281-297.

Meila, M. and D. Heckerman, 1998. An experimental comparison of several clustering and initialization methods. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, July 24-26, 1998, Morgan Kaufmann, San Francisco, CA., pp: 386-395.

Milligan, G.W. and M.C. Cooper, 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50: 159-179.

Milligan, G.W., 1981. A discussion of procedures for determining the number of clusters in a data set. Proceedings of the Classification Society Meeting, May 31-June 2, 1981, Toront, Canada.

Minaei-Bidgoli, B., A. Topchy and W.F. Punch, 2004. A comparison of resampling methods for clustering ensembles. Proceedings of the International Conference on Machine Learning: Models, Technologies and Application, June 21-24, 2004, Las Vegas, USA., pp: 939-945.

Mirkin, B., 2005. Clustering for Data Mining: A Data Recovery Approach. Chapman and Hall, London.

Modha, D.S. and W.S. Spangler, 2000. Clustering hyper text with applications to web searching. Proceedings of the 11th ACM on Hypertext and Hypermedia, May 30-June 3, 2000, San Antonio, TX., pp: 143-152.

Modha, D.S. and W.S. Spangler, 2003. Feature weighting in k-means clustering. Machine Learn., 52: 217-237.

Monti, S., P. Tamayo, J. Mesirov and T. Golub, 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learn., 52: 91-118.

Mufti, G.B., P. Bertrand and L.E. Moubarki, 2005. Determining the number of groups from measures of cluster stability. Proceedings of International Symposium on Applied Stochastic Models and Data Analysis, May 17-20, 2005, Brest, France, pp: 405-413.

Perruchet, C., 1983. Les epreuves de classifiabilite en analyses des donnees [Statistical tests of classifiability]. Tech. Rep. NT/PAA/ATR/MTI/810). C.N.E.T. Issy-Les-Moulineaux, France.

Selim, S.Z. and M.A. Ismail, 1984. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Trans. Pattern Anal. Mach. Intellg., 6: 81-87.

Shen, J., S.I. Chang, E.S. Lee, Y. Deng and S.J. Brown, 2005. Determination of cluster number in clustering microarray data. Applied Math. Comput., 169: 1172-1185.

Sneath, P.H.A. and R.R. Sokal, 1973. Numerical Taxonomy. W.H. Freeman and Company, San Francisco, USA., ISBN: 0-7167-0697-0.

Sugar, C.A., and G.M. James, 2003. Finding the number of clusters in a dataset: An information-theoretic approach. J. Am. Stat. Assoc., 98: 750-763.