Rishi Syal, Dr V.Vijaya Kumar / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 4, July-August 2012, pp.390-398

# Innovative Modified K-Mode Clustering Algorithm

***Rishi Syal, **Dr V.Vijaya Kumar**
*Prof and Head, Department of Computer Science and Engineering,
Guru Nanak Engineering College, AP (India)
**Dean and Professor, Department of CSE, IT and MCA, GIET
Rajahmundry, A.P., INDIA

## Abstract

**The K-Mode algorithm is known for applications on categorical datasets but with few drawbacks like selecting random k value, efficiency on cluster accuracy and so on. This paper provides research study on extension and modification of K-mode algorithm to provide good initial starting mean to get better clusters with better accuracy results on categorical data domains. The proposed algorithm has been experimented on large datasets with more than 2 lakh record and comparative study of traditional k-mode and proposed modified k-mode algorithm for varying data values has been shown for qualitative parameters.**

**Keywords:** Clustering, K-Mean, K-Mode, clustering accuracy

## I INTRODUCTION

Clustering can be recognized as the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [2]. Clustering approaches can be classified as partitioning [3-5], hierarchical [6] [7], density based [8] [9], fuzzy clustering [7], artificial neural clustering [10], statistical clustering, grid based, mixed and more [11]. In these approaches, Partitional and hierarchical clustering algorithms are two key approaches in research communities [2]. Partitional clustering aspires to directly acquire a single partition of the set of items into clusters. Most of these approaches are based on the iterative optimization of a criterion function depicting the "agreement" between the data and the partition [12]. Partitional clustering algorithms partition the data set into a particular number of clusters and then evaluate them on the basis of a criterion. These algorithms attempt to minimize particular criteria (for instance a square error function) and can be considered as optimization problems.

.

Partitioning methods are based on the principal that data are partitioned from the start into a fixed number of cluster and during the clustering process,
They make the change in the existing cluster(s) based on some similarity measure to the closest cluster. These algorithms typically allow the user to specify the number of clusters as an input parameter. Partitional algorithms can take numerical, categorical and mixed data to converge to some cluster.
Following concepts are opener to the various algorithms that are applied to the existing datasets.

## 2. Datasets

**2.1** Numerical data: The most extensively employed partitional algorithm is the iterative k-means approach. The k-means algorithm begins with k centroids (initial values are randomly chosen or derived from a priori information). Then, each pattern in the data set is allocated to the closest cluster (closest centroid). To conclude, the centroids are computed again as per the associated patterns. This procedure is done again until convergence is obtained [13]. Although K-means [5] was first introduced over 50 years ago, it is still regarded as one of the most extensively utilized algorithms for clustering. It is widely popular due to the ease of implementation, simplicity, efficiency, and empirical success [1].

K-Medoid or PAM(Partitioning Around Medoid) K-Medoid deals with the problem of outlier posed by K-means for ex if an object has very different value from the rest of values in a cluster , then that object will distort the result with k-mean because the mean value be changed to deal with the object . K-medoid is similar to K-mean except that mean of each cluster is the value which is nearest to the centre of cluster.

CLARA is an extension of K-medoid dealing with results of large datasets as K-medoid can't deal with the same.
Fuzzy K-means is another partitional algorithm that uses the concept of degree of membership in each cluster.

**2.2 Categorical data :** Categorical data also known as nominal or qualitative data is the core context of this research. Computing similarity between categorical data is not straightforward because there is no explicit notion of ordering between data. It is a type of data consisting of categories of data whose value is one of a fixed number category of data. They may be nominal categories which may have been

derived from observation made of qualitative data. It may consist of small number of values each corresponding to a specific category for ex mother employed (yes no), mother marital status (married, divorced, single, widowed), color of car (red, green, blue).

There have been many earlier versions of K-means like algorithm working on clustering categorical data. The k-mode algorithm extends the K-means algorithm by using simple matching dissimilarity for categorical objects, modes instead of means for clusters and frequency based method to update modes in the clustering process. To minimize cost functions, there are versions of k-means combinations with k-mode (Huang) resulting in the k-prototype algorithm for clustering objects that describe both numerical and categorical data. These algorithms have removed the drawback of numeric only limitation of the k-mean algorithm and enabled it to be used for efficient clustering of very large datasets from real datasets.

In real time applications, there is no clear boundary between clusters.

When fuzzy clustering is used, membership degrees between zero and one are used in fuzzy clustering instead of a very clear and crisp assignment. It is an extension of fuzzy K-means algorithm for clustering categorical data.

Squeezer was a one pass algorithm which repeatedly reads tuples from the dataset one by one. It takes the first tuple as it reads and creates cluster of it as alone and then the subsequent tuples are either put into existing cluster or rejected by all existing clusters to form new cluster by the given similarity function.

The k-means algorithm is a popular partitional algorithm for data clustering. K-means clustering is a method of cluster analysis which aims to partition $n$ observations into $k$ clusters in which each observation belongs to the cluster with the nearest mean. When we apply k-means algorithm to the categorical data, this has counteracted by two problems such as, (1) Formation of cluster center (2) Finding the similarity measure in between the cluster center and the categorical objects. By handling these two issues, we have used the k-mode algorithm [19] based on the k-means algorithm. K-mode assigns mode to each cluster as a summary to cluster's most frequent value attribute value. The algorithmic procedure of the k-means algorithm for categorical data (K-Mode algorithm) is discussed below.

## 3 K-means algorithm for categorical data (K-Mode algorithm)

Let us consider a categorical data set $D = \{X_1, \ldots, X_n\}$ of categorical objects to be clustered, where each object $X_i = (x_{i,1}, \ldots, x_{i,m}), 1 \le i \le n$ is defined by $m$ categorical attributes. Then, this problem can be mathematically formulated as follows:

Minimize

$$P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l} d(X_i, Q_l)$$

subject to, $\sum_{l=1}^{k} w_{i,l} = 1, \ 1 \le i \le n,$

$w_{i,l} \in \{0,1\}, \ 1 \le i \le n, 1 \le l \le k$

Where $W = [w_{i,l}]_{n \times k}$ is a partition matrix, $Q = \{Q_1, Q_2, \ldots, Q_k\}$ is a set of representatives, and $d(X_i, Q_l)$ is the dissimilarity between object $X_i$ and representative $Q_l$.

### 3.1 Basic steps of k-mode clustering algorithm:

1) Initialize $k$ representatives, one for each cluster.
2) Compute the dissimilarity $d(X_i, Q_l), \ l = 1,2,\ldots,k$ of each $k$ representative with categorical objects $X_i$ in $D$.
3) Assign categorical object $X_i$ to cluster $C_l$ whose dissimilarity measure is less.
4) Update the $k$ representatives based on definition 2.
5) Repeat Step 2 to step 4, until there is no movement of the objects between the clusters.

**Definition 1: (Dissimilarity Measure)**

Let X, Y be two categorical objects described by categorical attributes. The Dissimilarity measure between X and Y can be defined as total mismatches of the corresponding attribute categories of the two objects. The smaller the number of mismatches, the more similar the two objects X and Y are.

The dissimilarity of categorical object $X_i$ with the representative $Q_l$ is computed based on the following equations.

$$d(X_i, Q_l) = \sum_{j=1}^{m} \delta(x_{i,j}, q_{l,j})$$

$$\delta(x_{i,j}, q_{l,j}) = \begin{cases} 0; & if\ x_{i,j} = q_{l,j} \\ 1; & otherwise \end{cases}$$

**Definition 2: (Updating of k-representatives)**

Initially, the categorical object $X_i$ related with cluster $C_l$, $l = 1,2,\ldots,k$ are obtained and then, we compute the relative frequency $f_{x_{i,j}}$ of every category $x_{i,j}$ within the cluster $C_l$. The categories of categorical attributes within the cluster $C_l$ are arranged in accordance with their relative frequency $f_{x_{i,j}}$. The category $x_{i,j}$ with high relative frequency of '$m$' categorical attributes is chosen for the new representative. For example, gender is a categorical attribute having two categories (male and female) and hair color is also a categorical attribute having a number of categories (blonde, brown, brunette, red and more).

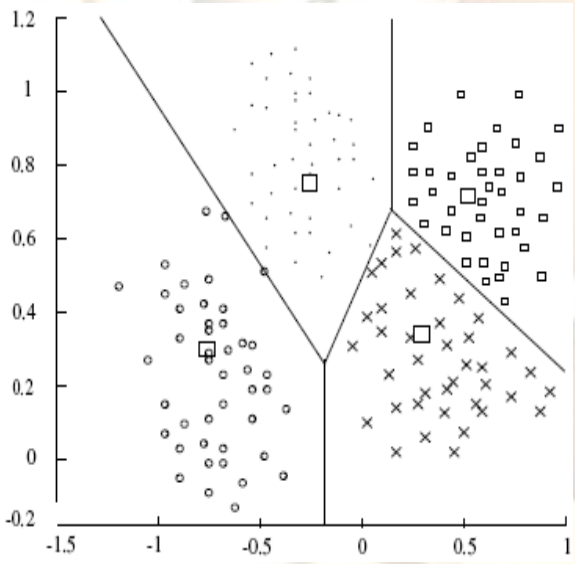


**Figure1:** Sample result of K-means clustering algorithm

K-mode has computation complexity of O (tkn) where n is the number of objects, k is the number of clusters and t is number of iteration. The result of K-mode is dependent on initial mode.

Zengyou He *et al*. [15] have analyzed clustering algorithms for categorical data on the basis of cross-fertilization between the two disjoint research fields. It was defined that the CDC (Categorical Data Clustering) problem was an optimization problem from the perspective of CE (Cluster Ensemble), and employed CE approach for clustering categorical data. The experimental results on real datasets demonstrated that CE based clustering method was competitive with available CDC algorithms with regard to clustering accuracy. Fuyuan Cao *et al*. [16] have described an initialization method for categorical data. It employed the initialization method to k-modes algorithm and fuzzy k-modes algorithm. Experimental results showed that the presented initialization method was better than random initialization method and can be employed to large data sets for its linear time complexity with regard to the number of data objects. Dae-Won Kim *et al*. [17] have extended k-modes-type algorithms for clustering categorical data by denoting the clusters of categorical data with k-populations as an alternative of the hard-type centroids employed in the conventional algorithms. Use of a population-based centroid representation enables it to maintain the uncertainty inherent in data sets as far as possible before actual decisions were finalized. The k-populations algorithm was noticed to provide better clustering results through various experiments.

## 4. Proposed Modified K_Mode algorithm

### Step a: Partitioning the Dataset into blocks

Partitioning is the first step of the proposed modified K-mode clustering algorithm. It is the process of partitioning the categorical dataset $D$ into $p$ blocks with equal size ($D_1, D_2, \cdots, D_p$). This partitioning process can be done randomly or systematically. Clustering across extremely large categorical data set, typically many million categorical objects, is a complex and time consuming process. Partitioning the categorical dataset can lead to better performance through parallel operations.

### Step b: Sub Clustering each block using modified K-Mode Algorithm

The $p$ blocks obtained from the previous step are then subjected to the modified k-mode algorithm for sub clustering. The final result of usual k-means clustering algorithm depends on the initial starting means. Therefore, we have devised an efficient method as in [23], to obtain good initial starting means of $p$ blocks, so that the final result is better than the randomly selected initial starting means. The algorithmic description of the modified k-means algorithm is given as follows:

1. The block ($D_i$) from the previous step is further partitioned into $m$ set of blocks and every

Block is subjected to the k-mode algorithm as discussed in previous section.

2. The obtained k means from every $m$ block are organized as a compact data of block $D_i$.
3. Again, we apply the k-mode algorithm in this compact data of $D_i$ and the final k- Mode gives the initial starting means for applying the k-mode algorithm over a block $D_i$. This provides $k$ number of sub clusters in each block so that we can obtain sub- clusters of size $p * k$.

## 5.0 EXPERIMENTAL ENVIRONMENT AND DATASETS

This experimental environment of proposed **K_Mode** clustering is Windows XP Operating system at a 2 GHz dual core PC machine with 2 GB main memory running a 64-bit version of Windows 2007. The experimental analysis of the proposed research methodology is presented in this section. For extensive analysis, I have utilized the very large categorical dataset taken from the Infobiotics PSP (Protein Structure Prediction) benchmarks repository.

***Dataset Description:*** The datasets utilized here for evaluation purposes is taken from Infobiotics PSP benchmarks repository. This repository contains an adjustable real-world family of benchmarks suitable for testing the scalability of classification/regression methods. When we test a machine learning method we usually choose a test suite containing datasets with a broad set of characteristics, as we are interested in checking how the learning method reacts to each different scenario. In this, we have chosen 4 different datasets (DB1, DB2, DB3 and DB4) which contains more than 2 lakh records. The dataset DB1 and DB4 contains 3 attributes and 2 class labels. As well, the datasets DB2 and DB3 contains 5 attributes and 2 class labels.

The results obtained from the experimentation of the proposed Modified K_Mode along with the traditional K_Mode are part of this section. This section deals with the comparative study of K_Mode and Modified K_Mode

### 5.1 EVALUATION METRICS

The obtained results from the proposed research are analyzed with the aid of the evaluation metrics such as clustering accuracy, clustering error, memory usage and the computational time. We have used the clustering accuracy as described in [19] for evaluating the performance of the proposed approach. The evaluation metric used in the proposed approach is given below,

$$\text{Clustering Accuracy, } CA = \frac{1}{N}\sum_{i=1}^{T} X_i$$

$$\text{Clustering Error, } \mathbf{CE = 1 - CA}$$

Where, $N$ → Number of data points in the dataset

$T$ → Number of resultant cluster

$X_i$ → Number of data points occurring in both cluster $i$ and its corresponding class.

Computational time indicates the time taken to execute the computer program and it typically grows with the input size. The Memory usage defines the memory utilized by the current jobs present in the particular system.

The experimental results have been evaluated for varying values of K on the traditional as well as proposed Modified K_Mode algorithm. .The study and analysis are described after the results section

### 5.2 Experimental Results

**Dataset TrainFold00w1.csv (db1)**

**K-Mode**
k=5
Memory Usage in Bytes: 40351424
Computation time in millisecond: 57797
Time taken=00:00:57:797
Clustering Accuracy=0.700685
Clustering Error=0.299315

**Modified K_Mode**
k=5
Memory Usage in Bytes: 48290136
Computation time in millisecond: 27109
Time taken=00:00:27:109
Clustering Accuracy=0.700685
Clustering Error=0.299315

---

**K-Mode**
k=6
Memory Usage in Bytes: 28462712
Computation time in millisecond: 43718
Time taken=00:00:43:718
Clustering Accuracy=0.700685
Clustering Error=0.299315

**Modified K_Mode**
k=6
Memory Usage in Bytes: 29878160
Computation time in millisecond: 23766
Time taken=00:00:23:766

Clustering Accuracy=0.700685
Clustering Error=0.299315

---

**K-Mode**
k=7
Memory Usage in Bytes: 35025144
Computation time in millisecond: 44734
Time taken=00:00:44:734
Clustering Accuracy=0.700685
Clustering Error=0.299315

**Modified k-Mode**
k=7
Memory Usage in Bytes: 39598016
Computation time in millisecond: 22656
Time taken=00:00:22:656
Clustering Accuracy=0.7006

---

**K-Mode**
k=8
Memory Usage in Bytes: 36226504
Computation time in millisecond: 19531
Time taken=00:00:19:531
Clustering Accuracy=0.700685
Clustering Error=0.299315

**Modified K_Mode**
k=8
Memory Usage in Bytes: 35281968
Computation time in millisecond: 17000
Time taken=00:00:17:00
Clustering Accuracy=0.700685
Clustering Error=0.299315

---

**TrainFold00w2.csv dataset (db2)**

---

**K_Mode**
K=5
Memory Usage in Bytes: 42011488
Computation time in millisecond: 63984
Time taken=00:01:03:984
Clustering Accuracy=0.53823
Clustering Error=0.46177

**Modified K_Mode**
**K=5**
Memory Usage in Bytes: 45159824
Computation time in millisecond: 31562
Time taken=00:00:31:562
Clustering Accuracy=0.529835
Clustering Error=0.470165

---

**K-Mode**
k=6
Memory Usage in Bytes: 46508208
Computation time in millisecond: 51547
Time taken=00:00:51:547
Clustering Accuracy=0.51849
Clustering Error=0.48151

**Modified K_Mode**
**K=6**
Memory Usage in Bytes: 46508208
Computation time in millisecond: 24347
Time taken=00:00:24:347
Clustering Accuracy=0.51849
Clustering Error=0.48131

---

**K_Mode**
k=7
Memory Usage in Bytes: 42068664
Computation time in millisecond: 39000
Time taken=00:00:39:00
Clustering Accuracy=0.533755
Clustering Error=0.466245
**Modified K_Mode**
k=7
Memory Usage in Bytes: 46508208
Computation time in millisecond: 25377
Time taken=00:00:25:37
Clustering Accuracy=0.54849
Clustering Error=0.44151

---

**K_Mode**
k=8
Memory Usage in Bytes: 39038904
Computation time in millisecond: 33766
Time taken=00:00:33:766
Clustering Accuracy=0.529315
Clustering Error=0.470685

**Modified K_Mode**
k=8
Memory Usage in Bytes: 41017056
Computation time in millisecond: 20656
Time taken=00:00:20:656
Clustering Accuracy=0.525005
Clustering Error=0.474995

---

**TrainFold03w2_org.csv dataset (db3)**

---

**K_Mode**
k=5
Memory Usage in Bytes: 40095784
Computation time in millisecond: 60375
Time taken=00:01:00:375
Clustering Accuracy=0.5277
Clustering Error=0.47230000000000005

**Modified K_Mode**
k=5
Memory Usage in Bytes: 45443912
Computation time in millisecond: 29562
Time taken=00:00:29:562
Clustering Accuracy=0.536
Clustering Error=0.46399999999999997

---

**K_Mode**
k=6
Memory Usage in Bytes: 42207800
Computation time in millisecond: 50344
Time taken=00:00:50:344
Clustering Accuracy=0.54247
Clustering Error=0.45753

**Modified K_Mode**
k=6
Memory Usage in Bytes: 43803896
Computation time in millisecond: 27563
Time taken=00:00:27:563
Clustering Accuracy=0.539555
Clustering Error=0.460445

---

**K_Mode**
k=7
Memory Usage in Bytes: 47972120
Computation time in millisecond: 41969
Time taken=00:00:41:969
Clustering Accuracy=0.53402
Clustering Error=0.46597999999999995

**Modified K_Mode**
k=7
Memory Usage in Bytes: 40361784
Computation time in millisecond: 23843
Time taken=00:00:23:843
Clustering Accuracy=0.533695
Clustering Error=0.466305

---

**K_Mode**
**k**=8
Memory Usage in Bytes: 43320384
Computation time in millisecond: 54328
Time taken=00:00:54:328
Clustering Accuracy=0.54675
Clustering Error=0.45325000000000004

**Modified K_Mode**
k=8
Memory Usage in Bytes: 39516072
Computation time in millisecond: 21562
Time taken=00:00:21:562
Clustering Accuracy=0.538885
Clustering Error=0.46111500000000005

---

Memory Usage in Bytes:39584976
Computation time in millisecond: 20032
Time taken=00:00:20:32
Clustering Accuracy=0.53584
Clustering Error=0.46416

---

**dataset: TrainFold04w1.csv (db4)**

---

**K_Mode**
---
k=5
Memory Usage in Bytes: 38995192
Computation time in millisecond: 30094
Time taken=00:00:30:94
Clustering Accuracy=0.530015
Clustering Error=0.469985

**Modified K_Mode**
k=5
Memory Usage in Bytes: 44559704
Computation time in millisecond: 28984
Time taken=00:00:28:984
Clustering Accuracy=0.53683
Clustering Error=0.46317

---

**K_Mode**
k=6

Memory Usage in Bytes: 40184560
Computation time in millisecond: 67734
Time taken=00:01:07:734
Clustering Accuracy=0.53228
Clustering Error=0.46772

**Modified K_Mode**
k=6
Memory Usage in Bytes: 45261672
Computation time in millisecond: 22703
Time taken=00:00:22:703
Clustering Accuracy=0.539975
Clustering Error=0.460025

---

**K_Mode**
k=7
Memory Usage in Bytes: 43604816
Computation time in millisecond: 41937
Time taken=00:00:41:937
Clustering Accuracy=0.54412
Clustering Error=0.45587999999999995

**Modified K_Mode**
k=7
Memory Usage in Bytes: 35992672
Computation time in millisecond: 21406
Time taken=00:00:21:406
Clustering Accuracy=0.546745
Clustering Error=0.453255

---

**K_Mode**
k=8
Memory Usage in Bytes: 41862192
Computation time in millisecond: 33203
Time taken=00:00:33:203
Clustering Accuracy=0.55071
Clustering Error=0.44928999999999997

**Modified K_Mode**
k=8
Memory Usage in Bytes: 30469592
Computation time in millisecond: 32343
Time taken=00:00:37:343
Clustering Accuracy=0.550185
Clustering Error=0.44981499999999996

---

**5.3 PERFORMANCE EVALUATION**

The performance of the proposed approach has been evaluated in terms of the evaluation metrics memory usage and the computational time. The metrics memory usage and the time have been analyzed with the four different very large categorical datasets by giving different *k* values and the thresholds.

**1) Performance Analysis of Proposed approach on DB1**

With the evaluated results of the accuracy and the error rate as shown with different *k* values, it shows that the proposed Modified k-mode algorithm yields better results in terms of accuracy and error rate when compared to the k-mode algorithm. As well, the results of the Modified k-mode algorithm are compared against with the k-mode algorithm by means of memory usage and the computational time by varying the k values. In this, the modified k-mode algorithm produce better results with lesser time and slightly more memory usage.

**2) Performance Analysis of Proposed approach on DB2**

The performance analysis results of the proposed approach on Dataset DB2 are presented in this section. The clustering accuracy and the error rate are evaluated for Modified k-mode algorithm, k-mode by analyzing the dataset DB2 with the evaluation metrics. By varying the *k* values, the evaluated results of the accuracy and the error rate shows that the Modified k-mode, k-mode. In addition, the results of the Modified k-mode algorithm are compared against with the k-mode algorithm by means of memory usage and the computational time by varying the k values. In this, the modified k-mode algorithm produce better results with lesser time and memory usage produce almost similar results for all the k values and slightly deviate in a few cases.

**3) Performance Analysis of Proposed approach on DB3**

The performance analysis results of the proposed approach on Dataset DB3 are presented in this section. While analyzing the dataset DB3 with the evaluation metrics, the clustering accuracy and the error rate is evaluated for Modified k-mode algorithm, K-Mode. With the evaluated results of the accuracy and the error rate shown respectively with different *k* values, it shows that the proposed Modified k-mode algorithm and the k-mode algorithm yields almost similar results in terms of accuracy and error rate . As well, the results of the Modified k-mode algorithm are compared against with the k-mode algorithm by means of memory usage and the computational time by varying the k values as shown above. In this, the modified k-mode

algorithm and the k-mode algorithm produce better results with time and memory usage

**4) Performance Analysis of Proposed approach on DB4**

The performance analysis results of the proposed approach on Dataset DB4 are presented in this section. While analyzing the dataset DB4 with the evaluation metrics, the clustering accuracy and the error rate is evaluated for Modified k-mode algorithm, K-Mode. With the evaluated results of the accuracy and the error rate as shown respectively with different *k* values, it shows that the proposed Modified k-mode algorithm yields better results in terms of accuracy and error rate when compared to the K-Mode. As well, the results of the Modified k-mode algorithm are compared against with the k-mode algorithm by means of memory usage and the computational time by varying the k values as shown above. In this, the modified k-mode algorithm produces better results with lesser time and slightly more memory usage.

## Conclusion

The biggest advantage of using K-Means algorithm is its efficiency in handling numeric data sets for clustering but when it comes to categorical data set, the proposed algorithm which is an extension of traditional K_Mode has been shown to work efficiently. The proposed Modified K_Mode algorithm has been tested against 4 sets of data with the parameters of accuracy, error rate, computation time and memory usage. It was found that the proposed K_Mode algorithm yields better results in terms of all the above said parameter particularly giving better computation time under different conditions set. The modified K_Mode has been tested for varying values of k to prove that this proposed approach works very well under varying conditions for a very large categorical data set. The same is true for even medium to smaller data sets. Thence this proposed approach is scalable to very large data sets.

The concept of modified K_Mode may works very well in hybrid clustering algorithm where the combined features of partitional and hierarchical algorithm works with the advantages of both of them being applied to get better efficient results while clustering.

## Acknowledgement

The authors would like to thank Sardar Tavinder Singh Kohli, Chairman, Sardar Gagandeep Singh Kohli Vice Chairman and Dr H. S. Saini Managing Director ,Guru Nanak Technical Institutions for their encouragement and support for this research. The authors would like to express their gratitude to the reviewers for their valuable suggestions and comments. The work is (partially) supported by research grants from the R&D of Guru Nanak Engineering College.

## References

[1] Anil K. Jain, "Data clustering: 50 years beyond K-means", Lecture Notes in Computer Science, Springer, vol. 5211, pp. 3-4, 2008.

[2] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[3] R.C. Dubes, "How Many Clusters Are Best?—An Experiment", Pattern Recognition, Vol. 20, No. 6, pp. 645-663, 1987.

[4] C.-R. Lin and M.-S. Chen, "On the Optimal Clustering of Sequential Data", In Proceedings of Second International Conference on Data Mining, April 2002.

[5] J. McQueen, "Some Methods for Classification and Analysis of Multivariate Observations", In Proceedings of Fifth Berkeley Symposium on Math.Statistics and Probability, Vol. 1, pp. 281-297, 1967.

[6] P.H.A. Sneath and R.R. Sokal,"Numerical Taxonomy. London: Freeman", 1973.

[7] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York: Plenum Press, 1981, ISBN: 0306406713.

[8] M.M. Breunig, H.-P. Kriegel, P. Kro¨ger, and J. Sander, "Data Bubbles: Quality Preserving Performance Boosting for Hierarchical Clustering", In Proceedings of ACM SIGMOD, Vol. 30, No. 2, pp. 79-90, 2001.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In Proceedings Second International Conference on Knowledge Discovery and Data Mining, pp. 226-231, 1996.

[10] J. Hertz, A. Krogh, and R.G. Palmer, "Introduction to the Theory of Neural Computation. Reading", Mass Addison-Wesley, 1991.

[11] Cheng-Fa Tsai and Chia-Chen Yen, "ANGEL: A New Effective and Efficient Hybrid Clustering Technique for Large Databases", Lecture Notes in Computer Science, Springer, pp. 817-824, 2007.

[12] Nizar Grira, Michel Crucianu, Nozha Boujemaa,"Unsupervised and Semi-supervised Clustering: a Brief Survey", Springer, August 16, 2005.

[13] F. Samadzadegan and S. Saeedi, "Clustering Of Lidar Data Using Particle Swarm Optimization Algorithm In Urban Area", 2009.
[14] Trevor Hastie, Robert Tibshirani and Jerome Friedman, "The Elements of statistical learning", Springer, Second Edition, 2008.
[15] Zengyou He, Xiaofei Xu and Shengchun Deng, "A cluster ensemble method for clustering categorical data", Information Fusion, Vol. 6, No. 2 , pp 143-151, June 2005.
[16] Fuyuan Cao,Jiye Liang and Liang Bai, "A new initialization method for categorical data clustering", Expert Systems with Applications, Vol. 36, No. 7, pp. 10223-102284, September 2009.
[17] Dae-Won Kim, KiYoung Lee, Doheon Lee, and Kwang H. Lee, "A k-populations algorithm for clustering categorical data", Pattern recognition, Vol. 38, No. 7, pp.1131-1134, July 2005.
[18] Swagatam Das, Ajith Abraham and Amit 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, January 2008.
[19] Zhexue Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, vol. 2, pp. 283–304, 1998.
[20] Shyam Boriah, Varun Chandola and Vipin Kumar, "Similarity Measures for Categorical Data: A Comparative Evaluation", In Proceedings of 2008 SIAM Data Mining Conference, Atlanta, GA, April 2008.
[21] D. W. Goodall, "A New Similarity Index Based on Probability", in Biometrics, vol. 22, pp. 882-907. 1966.
[22] N. M. Murty and G. Krishna, "A hybrid clustering procedure for concentric and chain-like clusters", International Journal of Computer and Information Sciences, vol. 10, no.6, pp.397-412, 1981.
[23] Moth'd Belal and Al-Daoud, "A New Algorithm for Cluster Initialization," World Academy of Science, Engineering and Technology (WASET), Vol. 4, pp. 74 -76, 2005.