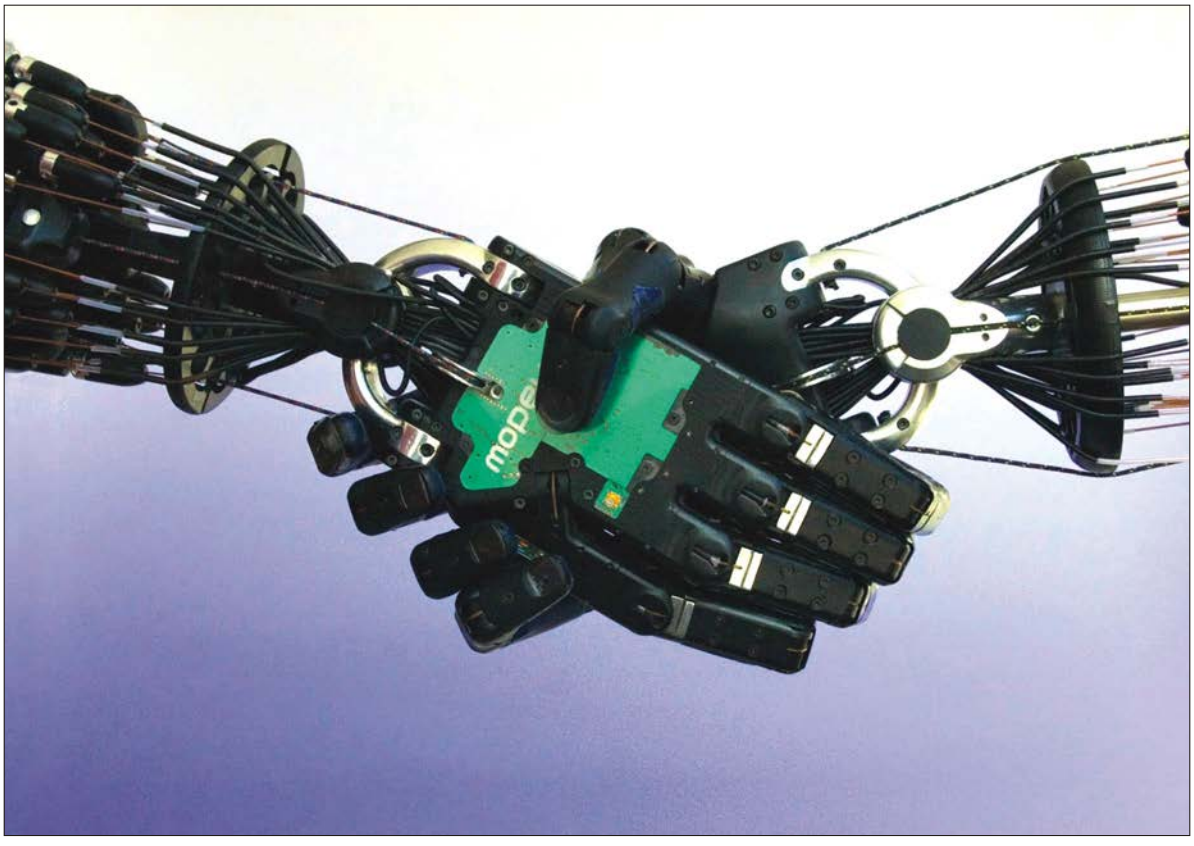

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Artificial Intelligence 6 (4): 257-265, 2013
ISSN 1994-5450 / DOI: 10.3923/jai.2013.257.265

A Medoid-based Method for Clustering Categorical Data

[1]Ali Seman, [1]Zainab Abu Bakar, [1]Azizian Mohd. Sapawi and [2]Ida Rosmini Othman
[1]Center for Computer Science Studies,
[2]Center for Statistical Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia

Corresponding Author: Ali Seman, Center for Computer Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Selangor, Malaysia Tel: +60355211191 Fax: +60355435100

ABSTRACT

Medoid-based method is an alternative technique to centroid-based method for partitional clustering algorithms. This method has been incorporated in a recently introduced clustering algorithm for categorical data, called k-Approximate Modal Haplotype (k-AMH) algorithm. This study reports the performance evaluation between the medoid-based method represented by the k-AMH algorithm and the centroid-based method represented by the extended k-Mode algorithm, the k-Population algorithm and the new Fuzzy k-Mode algorithm in clustering common categorical data. Nine common categorical data sets were used in the experiments to compare the performance of both methods using clustering accuracy scores. In overall results, the medoid-based method of k-AMH algorithm produced significant results for all data sets. The method showed its advantage of obtaining the highest clustering accuracy of 0.94 when clustering large number of clusters. This result indicated that the medoid-based method has a significant contribution for clustering categorical data, particularly for clustering large number of clusters.

Key words: Clustering algorithms, categorical data, partitional methods, centroid-based method, medoid-based method

INTRODUCTION

Clustering categorical data is one of the main clustering areas focused by many researchers. It begun when Huang (1998) published a new algorithm, called k-Modes algorithm designed specifically for categorical data. The k-Modes algorithm was introduced due to the ineffectiveness of k-Means algorithm (MacQueen, 1967) for clustering categorical data. Another effort, proposed by Ralambondrainy (1995) using a hybrid numeric-symbolic method also faced a problem of computational cost. As a result, the k-Modes algorithm has become the center of intention for solving categorical data.

A lot of variations of k-Modes-type algorithms have been extended and proposed for improving their clustering performances. A modification on the existing simple matching dissimilarity measure is one of the targets to improve the clustering results. For examples, He *et al.* (2007) have modified the original k-Modes algorithm (Later the algorithm is denoted as the Extended k-Modes algorithm) with four attribute weighting values called Relative Value Frequency (RVF), Uncommon Attribute Value Matches (UAVM), Hybrid method from RVF and UAVM (Here the method is denoted as Hybrid 1 or HI) and a simplified version of Hybrid I (Here the method is denoted as Hybrid II or HII). A similar idea of RVF method is also proposed by Ng *et al.*, (2007). The results reported by

He *et al.* (2007) show the improvement on the average clustering accuracy scores at almost greater than the average clustering accuracy scores of the original k-Modes algorithm for voting, breast cancer, mushroom, soybean, lymhography and zoo data sets.

On top of that, another significant result is made by Ng and Jing (2009). They proposed a new Fuzzy k-Modes algorithm with relative value frequency in fuzzy form and reported better clustering results for four categorical data sets: Soybean, Zoo, Credit and Adult. Furthermore, Kim *et al.* (2005) proposed a different idea through k-Population algorithm by introducing the soft centroids in the updating process of k-Population algorithm. No exact centroid is assigned explicitly like the other k-Mode-type algorithms. This idea has also produced impressive results in overall performances for categorical data sets such as Soybean, Zoo, Credit and Hepatitis.

Recently, a new algorithm, called k-Approximate Model Haplotype (k-AMH) algorithm has been proposed for clustering Y-Short Tandem Repeat (Y-STR) categorical data (Seman *et al.*, 2012). The algorithm was proposed due to the existing k-Mode-type algorithms could not handle the Y-STR data well. This is because the Y-STR data are unique and different with the other categorical data e.g., soybean, voting, mushroom, zoo, etc. The Y-STR data are composed of many similar and almost similar objects in inter or intra classes. This uniqueness of the Y-STR data has led the existing clustering algorithms to two problems. The first problem is the obtained centroids are not unique, thus resulting empty clusters. The second problem is the obtained centroids are insufficient to represent their clusters and therefore it leads to local minima problem. See the detailed description of the problems in clustering Y-STR data in Seman *et al.* (2012).

From observations during the experiments of clustering Y-STR data, it was found that the mode mechanism as the centroid is ineffective for high similarity data. The centroid-based method (the mode/mean mechanism) is always prone to bad initial center selection especially in handling high similarity data. The problem of initial center selections is typically the main issue, inherited from the k-Mean algorithm (Li *et al.*, 2008). As a solution for clustering Y-STR data, a medoid-based method has been introduced and associated in the k-AMH algorithm for solving the two problems above. As a result, the k-AMH algorithm has successfully produced better clustering results for clustering Y-STR categorical data e.g., the k-AMH algorithm recorded the highest mean accuracy score of 0.93 overall, compared to that of other algorithms: k-Population (0.91), k-Modes-RVF (0.81), New Fuzzy k-Modes (0.80), k-Modes (0.76), k-Modes-Hybrid 1 (0.76), k-Modes-Hybrid 2 (0.75), Fuzzy k-Modes (0.74) and k-Modes-UAVM (0.70). See the detailed results of the k-AMH algorithm in Seman *et al.* (2012).

This study reports the performance of the medoid-based of k-AMH algorithm against the other categorical data that were commonly used and experimented by the k-Mode-type algorithms e.g., the extended k-Modes algorithms (He *et al.*,2007), k-Population algorithm (Kim *et al.*, 2005) and New Fuzzy k-Modes algorithm (Ng and Jing, 2009). These selected algorithms were chosen for a comparison with the k-AMH algorithm because they had obtained impressive results when clustering several common categorical data sets.

METHODS

Classically, clustering is divided into hierarchical and non-hierarchical methods. The non-hierarchical method is also known as partitional method. The main difference between these two methods is the hierarchical method breaks up the data into a hierarchy of clusters, whereas the partitional method divides the data set into mutually disjoint partitions. The partitional methods

often adopt two popular heuristic methods: Centroid-based technique such as k-Means and Representative object-based technique (Medoid) such as k-Medoid algorithm (Han and Kamber, 2001; Tan *et al*., 2006).

Centroid-based methods: In centroid-based methods, a centroid of a cluster can be statistically represented by central of tendency measurements of data points referred as mean, mode and median. As a consequence, there are three categories of algorithms exist e.g., the k-Means-type algorithm, the k-Modes-type algorithm and the k-Median-type algorithm. However, the most well-known centroid-based technique is the k-Means algorithm (Han and Kamber, 2001). The k-Means algorithm begins in initializing cluster, k where k is a pre-defined number of clusters and uses the mean values to calculate the distance between objects and the k clusters. The distance measure is normally based on Euclidean distance. The algorithm allows recalculation of the means for each cluster of the objects that belong to it and minimizes the intra cluster dissimilarity. The updating centroid by the means is calculated as Eq. 1:

$$c_i = \frac{1}{m_i}\sum_{x \in l_i} x \tag{1}$$

where, c_i is the centroid cluster, l_i, x is an object and m_i is the number of objects in ith cluster.

When Huang (1998) proposed the k-Mode algorithm for clustering categorical data, he used a mode mechanism as the centroid. Thus, the updating centroids by mode is calculated as Eq. 2 and subject to (2a):

$$c_{ij} = a_j^{(r)} \tag{2}$$

where, $a_j^{(r)}$ is the mode of attribute values of A_j in cluster C_l such that:

$$f(a_j^{(r)} \mid c_l) \geq f(a_j^{(t)} \mid c_l), \forall t, 1 \leq t \leq p_j, a_j^{(r)} \neq a_j^{(t)} \tag{2a}$$

Medoid-based methods: Instead of using mean, mode and median, the medoid is based on object as the chosen center of a cluster. The most well known medoid method is the k-Medoid algorithm or Partitioning Around Medoids (PAM) introduced by Kaufman and Rousseeuw (1987). The algorithm plays around the objects which are the most centrally located object in a cluster. The basic idea of this algorithm is to find k cluster in n objects by first arbitrarily finding a representative object or often known as the medoid for each cluster. The next step is to iteratively replace one of the medoid by one of the non-medoid as long as the process can improve the clustering accuracy. The swapping technique allows exchange the current medoid, t_i and the non-medoid, t_h. The replacement of new medoid must satisfy the total cost, $TC_{ih}<0$ as Eq. 3:

$$TC_{ih} = \sum_{j=1}^{n} C_{jih} \tag{3}$$

where, C_{jih} is the cost change for an item t_j while swapping medoid, t_i with non-medoid, t_h.

However, the medoid mechanism used by the k-Medoid algorithm has faced a main problem of high computational cost (Ng and Han, 1994; Han *et al*., 2001; Dunham, 2003). As a consequence, many extended medoid-based methods have been proposed to improve the idea of k-Medoids algorithm e.g., Clustering Large Applications (CLARA) by Kaufman and Rousseeuw (1990) and Clustering Algorithm based on Randomized Search (CLARAN) by Ng and Han (1994). All efforts above are proposed for clustering numerical data. No effort of medoid-based method has been found for clustering categorical data except k-AMH algorithm.

k-AMH algorithm: The main differences between the k-AMH algorithm and the other k-Mode-type algorithms are (1) The objects (the data themselves) are used as the center (medoid) instead of modes and (2) Maximization process of the cost function is required instead of minimizing it as in the k-mode-type algorithms. The basic idea of the k-AMH algorithm is to find k clusters in n objects by first randomly selecting an object to be the medoid, h for each cluster. The next step is to iteratively replace the objects x one-by-one towards the medoid, h. The replacement is based on Eq. 4 if the cost function, P(À) as described in Eq. 5 and subject to Eq. 5a, 6, 6a, b, 7, 3, 4a, 7b and 7c is maximized:

$$P(\grave{A})^{r} > P(\grave{A})^{t}, r \neq t; \forall t, 1 \leq t \leq (n-k) \tag{4}$$

The P(À) is a cost function as described in Eq. 5:

$$P(\grave{A}) = \sum_{l=1}^{k}\sum_{i=1}^{n}\grave{A}_{li} \tag{5}$$

Subject to:

$$\grave{A}_{li} = W_{li}^{\alpha} D_{li} \tag{5a}$$

where, W^{α}_{li} is a (k×n) partition matrix that denotes the degree of membership of object i in the lth cluster that contains a value of 0-1 as described in Eq. 6, subject to Eq. 6a and b:

$$w_{li}^{\alpha} = \left\{ \begin{array}{ll} 1, & \text{If, } X_i = H_l \\ 0, & \text{If, } X_i = H_z,\ z \neq l \\ \left[\sum_{z=1}^{k} \left[\frac{d(X_i, H_l)}{d(X_i, H_z)} \right]^{1/(\alpha-1)} \right]^{-1}, & \text{If } H_i \neq X_j \text{ and } X_i \neq H_z,\ 1 \leq z \leq k \end{array} \right\}^{\alpha} \tag{6}$$

subject to:

$$w^{\alpha}_{li} \in [0, 1],\ 1 \leq i \leq n,\ 1 \leq l \leq k \tag{6a}$$

and:

$$0 < \sum_{i=1}^{n} w_{li}^{\alpha} < n, 1 \le l \le k \tag{6b}$$

Where:

- k (≤ n) is a known number of clusters
- H is the medoid such that $[H_1, H_2, \ldots, H_k] \epsilon X$
- $\alpha \epsilon [1, \infty)$ is a weighting exponent. Note that this alpha is typical based on 1.1 until 2.0 as introduced by Huang and Ng (1999)
- $d(X_i, H_z)$ is the distance measure between the object X_i and the medoid, H_l

D_{li} is another (k×n) partition matrix in which d_{li} contains a dominant weighting value of 1.0 or 0.5. The dominant weighting values are based on the value of w_{li}^{α} above. D_{li} is described in Eq. 7, subject to Eq. 7a-c:

$$d_{li} = \begin{cases} 1.0, & \text{if } w_{li}^{\alpha} = \max^{w_{li}^{\alpha}, 1 \le l \le k} \\ 0.5, & \text{otherwise} \end{cases} \tag{7}$$

subject to:

$$d_{li} \epsilon \{1.0, 0.5\}, 1 \le i \le n, 1 \le l \le k \tag{7a}$$

$$1.5 \le \sum_{l=1}^{k} d_{li} \le k, 1 \le i \le n \tag{7b}$$

$$0.5 < \sum_{i=1}^{n} d_{li} < n, 1 \le l \le k \tag{7c}$$

The detailed descriptions and the proof of the convergence of the medoid-based of k-AMH algorithm have been provided in Seman *et al*. (2012).

RESULTS AND DISCUSSION

Evaluation methods: Since, the results of the chosen k-Mode-type algorithms have previously been reported and published; the results and discussions here are only restricted to compare the k-AMH algorithm and the reported results as follows:

- The results obtained by the extended k-Modes algorithm as reported in He *et al*. (2007) for Voting, Breast Cancer, Mushroom, Soybean, Lymhography and Zoo data sets
- The results obtained by the k-Population algorithm as reported in Kim *et al*. (2005) for Soybean, Zoo, Credit and Hepatitis data sets
- The results obtained by the New Fuzzy k-Modes algorithm as reported in Ng and Jing (2009) for Soybean, Zoo, Credit and Adult data sets

This evaluation is based on the average clustering accuracy scores obtained from 100 experiments for each data set. Thus, the misclassification matrix proposed by Huang (1998) was used for obtaining the clustering accuracy scores and the performance of the algorithms was measured by the clustering accuracy, r defined by Huang (1998) as described in Eq. 8:

Table 1: Average clustering accuracy scores between the k-AMH algorithm and the extended k-modes algorithms

		Extended k-modes				
Data set	No. of class	RVF[a]	UAVM[b]	HI[c]	HII[d]	k-AMH algorithm
Voting	2	0.86	0.87	0.87	0.87	0.88
Breast cancer	2	0.86	0.87	0.91	0.95	0.90
Mushroom	2	0.75	0.76	0.73	0.73	0.88
Soybean	4	0.89	0.86	0.93	0.94	0.98
Lymphography	4	0.70	0.69	0.72	0.72	0.72
Zoo	7	0.87	0.82	0.86	0.85	0.94

RVF: Relative value frequency, UAVM: Uncommon attribute value matches, HI: Hybrid I-a combination method of RVF and UAVM and HII: Hybrid II-a simplified version of hybrid I

Table 2: Average clustering accuracy scores between the k-AMH algorithm and the k-population algorithm

Data set	No.of class	k-population	k-AMH algorithm
Soybean	4	1.00	0.98
Zoo	7	0.87	0.94
Credit	2	0.84	0.69
Hepatitis	2	0.80	0.79

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \quad (8)$$

where k, is the number of clusters, a_i is the number of instances occurring in both cluster i and its corresponding class and n is the number of instances in the data sets.

k-AMH vs extended k-mode algorithms: Table 1 shows a comparison result between the k-AMH algorithm and the extended k-Modes algorithm. This result was based on six categorical data sets: Voting, Breast Cancer, Mushroom, Soybean, Lymphography and Zoo. However, most of the data sets are 2 classes only. In overall, the k-AMH algorithm performs better than the extended k-Modes algorithm. The k-AMH algorithm obtained the highest clustering accuracy scores for four out of six data sets and produced an equal performance for a data set called, Lymphography. However,the k-AMH algorithm produced a bit lower of clustering accuracy score for a data set called, Breast cancer. The k-AMH algorithm also looks better performance when clustering a data set with large number of cluster e.g., Zoo data set (Seven clusters). A conclusion can be made; the medoid-based method implemented by the k-AMH algorithm has shown impressive performance over the data sets.

K-AMH vs k-population algorithms: Table 2 shows a comparison result between the k-AMH algorithm and the k-Population algorithm. This result was based on four categorical data sets: Soybean, Zoo, Credit and Hepatitis data sets. In overall, the k-Population algorithm performs better than the k-AMH algorithm. The k-AMH algorithm only obtained the highest clustering accuracy scores for one out of four data sets. The highest score obtained by k-AMH algorithm is 0.94 compared to only 0.87 obtained by k-Population algorithm for Zoo data set. This result clearly indicates that the k-AMH algorithm can perform better for the large number of cluster of the data set. In contrast, the algorithm faced a problem to cluster the data sets with two clusters/classes such as Credit and Hepatitis data sets.

K-AMH vs new k-mode algorithms: Table 3 shows a comparison result between the k-AMH algorithm and the New Fuzzy k-Modes algorithm. This result was based on four categorical data sets: Soybean, Zoo, Credit and Adult data sets. In overall, the New Fuzzy k-Modes algorithm performs better than the k-AMH algorithm. The k-AMH algorithm only obtained the highest clustering accuracy scores for one out of four data sets. The highest score is also for Zoo data set (0.94) compared to only 0.82 obtained by the New Fuzzy k-Modes algorithm. For Soybean data set, the New Fuzzy k-Modes algorithm also obtained the optimum score of 100% of clustering accuracy score; the same performance as obtained by the k-Population algorithm.

Big O comparison: The algorithm performances can also be evaluated in terms of the Big O analysis. It is to measure the efficiency of the execution time for the algorithm to run as a function of the input size. For clustering categorical data, usually the time efficiency of the algorithms may involve the following factors:

- The number of clusters, k
- The number of attributes, m
- The number of categories, j
- The number of iterations, t
- The number of objects, n

Table 4 shows a comparison of clustering categorical algorithms including the k-AMH algorithm. It shows that the k-AMH algorithm required only three factors which are k, m and n

Table 3: Average clustering accuracy scores between the k-AMH algorithm and the new fuzzy k-modes algorithm

Data set	No. of class	New fuzzy k-modes	k-AMH algorithm
Soybean	4	1.00	0.98
Zoo	7	0.82	0.94
Credit	2	0.80	0.69
Adult	2	0.75	0.73

Table 4: Big O comparison between the k-AMH algorithm and the other algorithms

Algorithm	Big-O
k-AMH	O(km(n-k)), where k is the No. of clusters, n is the No. of data and m is the No. of attributes
k-population	O(kpn(N+1)s), where k is the No. of clusters, p is the No. of attributes, n is the No. of data, N(=max(n$_j$)) is the maximum No. of categories for $1 \le j \le p$ and s is the No. of iterations required (Kim *et al*., 2005)
New fuzzy k-modes	No computational cost was given in Ng and Jing (2009). However, the cost is almost similar to the fuzzy k-modes algorithm or the k-modes-RVF algorithm
Fuzzy k-modes	O(kn(m+M), where k is the No. of clusters, m is the No. of attributes: $M\left(=\sum_{j=1}^{m} n_j\right)$ is the total No. of categories and n is the No. of objects (Huang and Ng, 1999)
k-modes	No computational cost was given in Huang (1998). However, the cost is almost similar to the fuzzy k-modes algorithm above
The extended k-modes with RVF, UAVM, HI and HII methods	O(tmnk), where t is the total No. of iteration required, m is the No. of attributes, n is the No. of objects and k is the No. of clusters (He *et al*., 2007). Note that this computational cost may differ from each other due to the complexity of that particular weighting schema. The most consumed time are the HI and UAVM methods

without factors j and t in its clustering process. The absence of j is because it uses the medoid-based method. Furthermore, the t factor is not required because the iteration is fixed from (n-k). It clearly indicates that the time complexity of k-AMH algorithm is still linear. Therefore, the algorithm is acceptable for clustering any categorical data.

CONCLUSION

From the experimental results, in general the k-AMH algorithm can be used for clustering any categorical data. Furthermore, the algorithm shows its better clustering performance when dealing with a large cluster number. This scenario can be seen from the result of Zoo data set with 7 clusters. In fact, the k-AMH algorithm has produced the best clustering accuracy scores over to the other two algorithms. This performance of the algorithm has been proven for clustering Y-STR categorical data. The k-AMH algorithm produced the highest clustering accuracy scores even though for clustering eight and 14 clusters of Y-STR data sets over the other eight clustering algorithms. In addition, the k-AMH algorithm is also efficient for clustering categorical data as its time complexity is only O(km(n-k)). As a conclusion, the k-AMH algorithm has its significant contribution on clustering any categorical data.

ACKNOWLEDGMENT

This research was supported by Fundamental Research Grant Scheme, Ministry of Higher Education, Malaysia. We would like to thank RMI, UiTM for their support for this research. We would like to extend our gratitude to many contributors toward the completion of this study our research assistants; Syahrul, Azhari, Kamal, Hasmarina, Nurin, Soleha, Mastura, Fadzila, Syukriah and Hazira.

REFERENCES

Dunham, M., 2003. Data Mining: Introductory and Advanced Topics. Prentice Hall, USA.

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, CA.

Han, J., M. Kamber and A.K.H. Tung, 2001. Spatial Clustering Methods in Data Mining: A Survey. In: Geographic Data Mining and Knowledge Discovery, Miller, H. and J. Han (Eds.). Taylor and Francis, USA., pp: 1-29.

He, Z., X. Xu and S. Deng, 2007. Attribute value weighting in K-modes clustering. Report Number Tr-06-0615, Cornell University Library, Cornell University, Ithaca, NY., USA. http://arxiv.org/abs/cs/0701013.

Huang, Z. and M.K. Ng, 1999. A Fuzzy k-Modes algorithm for clustering categorical data. IEEE Trans. Fuzzy Syst., 7: 446-452.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery, 2: 283-304.

Kaufman, J. and P.J. Rousseeuw, 1987. Clustering by Means of Medoid. In: Statistical Data Analysis Based on the L_1 Norm, Dodge, Y. (Ed.). Elsevier/North-Holland, Amsterdam, The Netherland, pp: 405-416.

Kaufman, L. and P.J. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York..

Kim, D.W., K.Y. Lee, D. Lee and K.H. Lee, 2005. A k-populations algorithm for clustering categorical data. Pattern Recognition, 38: 1131-1134.

Li, M.J., M.K. Ng, Y.M. Cheung and J.Z. Huang, 2008. Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters. Knowl. Data Eng., 20: 1519-1534.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, January 17-20, 1967, Berkeley, CA., USA., pp: 281-297.

Ng, R. and J. Han, 1994. Efficient and effective clustering methods for spatial data mining. Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, 1994, San Francisco, CA., USA., pp: 144-155.

Ng, M.K., M.J. Li, J.Z. Huang and Z. He, 2007. On the impact of dissimilarity measure in k-Modes clustering Algorithm. Trans. Pattern Anal. Mach. Intell., 29: 503-507.

Ng, M.K. and L. Jing, 2009. A new fuzzy k-Modes clustering algorithm for categorical data. Int. J. Granular Comp. Rough Sets Intell. Syst., 1: 105-119.

Ralambondrainy, H., 1995. A conceptual version of the k-Means algorithm. Pattern Recognit. Lett., 16: 1147-1157.

Seman, A., Z.A. Bakar and M.N. Isa, 2012. An efficient clustering algorithm for partitioning Y-short tandem repeats data. BMC Research Notes, Vol. 5. 10.1186/1756-0500-5-557

Tan, P.N., M. Steibach and V. Kumar, 2006. Introduction to Data Mining. Pearson Addison Wesley, Boston, MA., USA., ISBN-13: 9780321420527, Pages: 769.