Research Journal of Applied Sciences, Engineering and Technology 7(5): 963-969, 2014
DOI:10.19026/rjaset.7.342
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 29, 2013 Accepted: March 14, 2013 Published: February 05, 2014

Research Article

Performance Comparison of Clustering Techniques

Sambourou Massinanke and Lu Zhimao
College of Information and Communication Engineering, Harbin Engineering University, Harbin, China

Abstract: Data mining consists to extracting or "mining" information from large quantity of data. Clustering is one of the most significant research areas in the domain of data mining. Clustering signifies making groups of objects founded on their features where the objects of the same groups are similar and those belonging in different groups are not similar. This study reviews two Clustering Algorithms of the representative clustering techniques: K-modes and K-medoids algorithms. The two algorithms are experimented and evaluated on partitioning Y-STR data. All these algorithms are compared according to the following factors: certain number times of run, precision and recall. The global results show that K-mode clustering is better than the k-medoid in clustering Y-STR data.

Keywords: Data clustering, k-medoids clustering and data of Y-STR, k-modes clustering

INTRODUCTION

Clustering can be regarded as the most significant unsupervised learning issue; so, as every other problem of this kind, it consists to find a structure in a set of unlabeled data (Jain and Dubes, 1988; Jain *et al*., 1999). The K-means algorithm is irritable to outliers because of an object with a high value may greatly damage the distribution of data. How the algorithm can be modified to decrease this sensitivity? In place of taking the mean value of the instances in a cluster as a reference point, a Medoid can be utilized, that's the most centrally located instance in a cluster. Therefeore the partitioning method can be executed based on the principle of minimizing the sum of the dissimilarities between each instance and its corresponding reference point; this constitutes the idea of the K-Medoids method. The main strategy of K-Mediods algorithms is to determine K clusters in n objects by randomly finding a representative object (the Medoids) for all the clusters, each object is clustered with the Medoid to which it is the most closer. K-Medoids method utilizes representative observations as reference points in place of taking the mean value of the observations in each cluster. The algorithm assumes the input parameter K, the number of clusters that will contain n objects.

Clustering categorical data is a substantial research item in data mining. The *K*-modes algorithm (Huang, 1998) expands the *K*-means model to cluster categorical data. Since the *K*-modes algorithm applies the same clustering process as *K*-means, it conserves the effectiveness of the *K*-means algorithm. Presently, some *K*-modes based on clustering algorithms have been suggested (He *et al*., 2005; Gan *et al*., 2005).

Although the *K*-modes algorithm is remarkably efficacious technique, it shows two famous lacks as *K*-means algorithm:

- The solutions are just locally optimal.
- Their accuracies are sensitive to the initial conditions.

To surmount locally optimal in *K*-modes clustering, some algorithms such as tabu search (Ng and Wong, 2002) and genetic algorithm (Gan *et al*., 2005) have been proposed to determine the globally optimal solution. But, they are not able to afford approximation guarantees. Therefore, efficient approximation techniques would be elaborated for *K*-modes clustering. For our best knowledge, such kinds of approximation techniques are still not disposable nowadays.

Partitioning algorithms deals with K-means (Hartigan and Wong, 1979) and K-medoids (Kaufman and Rousseeuw, 2005). A description of these and other clustering techniques are studied in Jain *et al*. (1999). The K-means algorithm is the most famous among these algorithms due to its effectiveness and simplicity. The K-medoids algorithms have been proved to be more robust because they are less sensitive to the outliers and don't show limitations on attribute types while K-means are limited to multi-dimensional continuous datasets; and also, the clustering found is independent of the input order of the dataset. Furthermore, they remain invariant to orthogonal transformations and translations of the observations (Kaufman and Rousseeuw, 2005)

Corresponding Author: Sambourou Massinanke, College of Information and Communication Engineering, Harbin Engineering University, Harbin, China

The drawback of the K-medoids based algorithms is the time consuming thereby, they cannot be applied to large datasets. This encouraged the research of many approaches aiming to reduce the computational effort required to run these algorithms (Ester *et al*., 1995; Chu *et al*., 2002; Zhang and Couloigner, 2005).

A general strategy is to sample and apply the clustering algorithm to the resulting subset of objects. The accuracy of the resulting clusters is generally dependent on the selection of a relevant subset of objects.

The problem of clustering in general consists to partition a dataset consisting of n points (in m-dimensional space) into K distinct clusters in such way that the data objects in the same cluster are more close to each other than to objects in other clusters. The three questions (Ahmad and Dey, 2007) related to the clustering process are:

- Determining a similarity (or distance) between different data objects
- Executing an efficacious algorithm to find the clusters of most similar objects in an unsupervised way.
- Deduce a description that can distinguish the objects of a cluster in a brief way.

Classic clustering algorithms utilize Euclidean distance measure to estimate the similarity of two data objects (Haung *et al*., 2005; Krishna and Murty, 1999). That gives good results if the attributes of a dataset are simply numeric in nature. Nevertheless, Euclidean distance measure collapses to determine the similarity of data objects if the attributes are categorical or mixed. The data mining community is submerged with high collection of categorical data (Jain *et al*., 1999) such as these retrieved from health sectors, banks and biological data. The sector of Banking or the sector of health data are mainly combined data containing numeric attributes like salary, age, etc. and categorical attributes such as: sex, smoking or non-smoking, etc. Clustering combined datasets into significative groups is a challenging problem in which a good distance measure that can sufficiently determine data similarities (Chaturvedi *et al*., 2001). For handling mixed numeric and categorical data, some of the methods that were employed are as follows:

- Other method has been to discretize numeric attributes and apply categorical clustering algorithm. However, the discretization process conducts to loss information.
- The numeric distance can be used for calculating similarity between object pairs after transformation of nominal and categorical attribute values to numeric integer values. Nevertheless, it is very hard to give correct numeric values to categorical values.

The computational complexity of the PAM (Partitioning around Medoids) algorithm encouraged the development of CLARA (Clustering LARge Applications), a K-medoid algorithm based on sampling. CLARA utilizes many samples of the dataset and uses PAM on each one. Therefore, it chooses the clusters obtained from the execution, which gave the lowest objective function value and assigns each object of the entire data to the corresponding medoids. Kaufman and Rousseeuw (2005) shows that five samples of size 40+2K give satisfactory results. O (p^2/ K + K(p- K)) (Where p>K is the size of the sample) is the computational complexity of each iteration of CLARA to process each sample, thus it is faster than PAM. CLARANS was designed to enhance CLARA. It applies a randomized search strategy in order to enhance both Partitioning Around Medoids and Clustering LARge Applications algorithms in terms of efficiency (computation complexity/time) and effectiveness (average distortion over the distances) respectively.

The first element chosen is the object that has the minimum sum of dissimilarities (distances) to every other element (the objective function), so, the first element chosen is the dataset medoid. The other (K-1) medoids are chosen, one at a time, considering the elements that most reduce the objective function. When looking for new good medoids, CLARANS at random selects elements from the rest (n-K) elements, searching for the medoids of each group as its group center. The number of elements attempted in this step is limited by a user-provided parameter (maxNeighbor).

Despite the success of Sun *et al*. (2002), the following observations encourage us to continue other alternative initialization methods:

- The clustering resulting of iterative initial-points refinement algorithm seen in He *et al*. (2005) is random in nature. Then, different executions of the algorithm lead to different clustering results. To find clear clustering output, the end-user still must execute the algorithm repeatedly.
- He *et al*. (2005) shows (in experimental results) that very poor clustering results can occur in some cases. Therefore, non-randomized initialization algorithm is needed in real applications.
- Simple and easy to implement should be the new initialization method. It is expected that such initialization method merits good scalability.
- It would be very advantageous if the new initialization algorithm can furnish performance guarantee to certain degree.

Concerning the *K*-modes algorithm, a lot of research (Huang and Ng, 1999; Kim *et al*., 2004) have been led to enhance its performance. Huang and Ng present the Fuzzy *K*-modes algorithm (Huang and Ng,

1999) which assigns membership degrees to data objects in different clusters, the technique is applied in Ng *et al*. (2002) and genetic algorithm is used in Gan *et al*. (2005) to enhance *K*-modes algorithm. Optionally, fuzzy *k*-modes algorithm is expanded by representing the clusters of categorical data with fuzzy centroids in place the hard-type centroids used in the classic algorithm (Kim *et al*., 2005; Kim *et al*., 2004), but, most of these methods are slower than the classic *K*-modes algorithm in running time. Because of the *K*-modes algorithm is sensitive to the initial conditions, another realizable way for enhancing its performance is to elaborate efficient initialization methods. Finally, an iterative initial-points improvement algorithm for *K*-modes clustering is introduced in He *et al*. (2005).

K-MEDOID CLUSTERING

The K-Medoids algorithm was first introduced in Kaufmann and Rousseeuw (1990) and is not as sensitive to outliers as is the K-means. In this algorithm, each cluster is represented by the most centrally located object known as medoid.
The general procedure for the algorithm is as follows:

- Randomly choose K objects as the initial medoids.
- Assign each one of the remaining objects to the cluster that has the closest medoid.
- In a cluster, randomly select a nonmedoid object, which will be referred to as $O_{nonmedoid}$
- Compute the cost of replacing the medoid with $O_{nonmedoid}$ .this cost is the difference in the square error if the current medoid is replaced by $O_{nonmedoid}$. If it is negative, then make $O_{nonmedoid}$ the medoid of the cluster. The square error is again the summed error of all objects in the database:

$$E = \sum_{i=1}^{K} \sum_{o \in C_i} \left| o - O_{medoid(i)} \right|^2$$

where, $O_{medoid(i)}$ is the medoid of the i^th^ cluster.

- Repeat from (2) until there is no change.

K-MODE CLUSTERING

Most clustering algorithms focused on numerical dataset (Chaturvedi *et al*., 2001). However, much of the data existed in the databases is categorical, where attribute values cannot be naturally ordered as numerical values.

Various clustering algorithms have been reported to cluster categorical data. He *et al*. (2005) proposed a cluster ensemble for clustering categorical data. Ralambondrainy (1995) presented an approach by using k-means algorithm to cluster categorical data. The approach is to convert multiple category attributes into binary attributes (using 0 and 1 to represent either a category absent or present) and treat the binary attributes as numeric in the k-means algorithm. Gowda and Diday (1991) used other dissimilarity measures based on “position”, “span” and “content” to process data with categorical attributes. Huang (1998) proposed K-modes clustering which extend the k-means algorithm to cluster categorical data by using a simple matching dissimilarity measure for categorical objects. Recently, (Chaturvedi *et al*., 2001) also presented K-modes which used a nonparametric approach to derive clusters from categorical data using a new clustering procedure. Huang (2003) has demonstrated the equivalence of the two independently developed K-modes algorithm given in two papers which done by Huang (1998) and Chaturvedi *et al*. (2001). Then, San *et al*. (2004) proposed an alternative extension of the K-means algorithm for clustering categorical data which called K-representative clustering.

In this study, we concern to adopt K-mode clustering algorithm which was proposed by Huang (1998). This method is based on K-means clustering but remove the numeric data limitation. The modification of K-means algorithm to k-modes algorithm as follows (Huang, 1998).

- Using a simple matching dissimilarity measure for categorical objects
- Replacing means of clusters by mode
- Using a frequency based method to update the modes

The simple matching dissimilarity measure can be defined as following. Let X and Yare two categorical objects described by m categorical attributes. The dissimilarity measure between X and Y can be defined by the total mismatches of the corresponding attribute categories of the two objects. The smaller the number of mismatches is, the more similar he two objects. Mathematically, it can be represented as follows (Gowda and Diday, 1991):

$$d\left(X,Y\right) = \sum_{j=1}^{m} \delta\left(x_j, y_j\right) \tag{1}$$

where,

$$\delta\left(x_j, y_j\right) = \begin{cases} 0\, if\ x_j = y_j \\ 1\, if\ x_j \neq y_j \end{cases} \tag{2}$$

When (1) is used as the dissimilarity measure for categorical objects, the cost function becomes:

$$p\left(W,Q\right) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{i,l}\, \delta\left(x_{i,j}, x_{l,j}\right) \tag{3}$$

where, $w_{i,l} \in W$ and $Q_l = \left[q_{l,1}, q_{l,2}, \ldots, q_{l,m}\right] \in Q$

The k-modes algorithm minimizes the cost function defined in Eq. (3). The k modes algorithm consists of the following steps (Huang, 1998):

- Select K initial modes, one for each cluster.
- Allocate an object to the cluster whose mode is the nearest to it according to (1).
- After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters.
- Repeat 3 until no object has changed clusters after a full cycle test of the whole dataset.

Y-STR DATA AND ITS APPLICATIONS

Y-STR is defined as: Short Tandem Repeats on Y-Chromosome, Y-STR data expresses the number of times an STR repeats, called allele value for each marker. This DNA method is nowadays very used in Anthropological Genetics as well as in Genetic Genealogy. Moreover, this method is a very hopefully method to sustain a traditional approach especially in studying human migration patterns and proving genealogical relationships. For further information, the Y-STR used in Anthropology can be found in a book called Anthropological Genetics: Theory, Methods and Applications (2007) and for Genetic Genealogy can be found in Fitzpatrick (2005) and Fitzpatrick and Yeiser (2005). The genetic distance for a person may differ from other by referring the allele values for each marker. If a person shares the same allele value for each marker is considered coming from the same ancestor from genealogical perspective. In a broader perspective, for instance in studying human migration patterns, it can be under the same haplogroups (In molecular evolution, a haplogroup is a group of similar haplotypes that share a common ancestor having the same Single Nucleotide Polymorphism (SNP) mutation in both haplotypes. Because a haplogroup consists of similar haplotypes, this is what makes it possible to predict a haplogroup from haplotypes) which includes different geographical X area throughout the world. The Y-STR data can be grouped into meaningful groups based on the distance for each STR marker. For genealogical data such as Y-Surname project, the distances are based on 0 or 1 or 2 or 3 mismatches, whereas the haplogroups are determined by a method known as Single Nucleotide Polymorphism (SNP) analysis. There are set of very broad haplogroups and all males in the world can be placed into a system of defining Y-DNA haplogroups by letters A through to T, with further subdivisions using numbers and lower case letters. See International Society of Genetic Genealogy (www.isogg.org). The haplogroups have been established by the Y Chromosome Consortium (YCC). For further details, see University of Arizona (http://ycc.biosci.arizona.edu/).

NOTATIONS

Let X = {X_1, …, X_n} be set of n Y-STR data and XA = {A_1, …, A_n} bet set of markers/attributes of Y-STR. We define A_j is the j-the attributes values as associated j-th marker with the actual STR allele value. We define X is a numerical data if it is treated only as numerical values as it is. Note that the Y-STR data are originally a numeric domain as associated with the allele values and it is discrete values. We define X is a categorical data if it is treated only as categorical values. Note that for each attribute A_j describes a domain values, denoted DOM (A_j). A domain DOM (A_j) is defined as categorical data if it is finite and unordered, e.g., for an a, b ∈ DOM (A_j) either a = b or $a \neq$ b. Consider the j-th attribute values are: A_j = {10, 10, 11, 11, 12, 13, 14}, thus the Dom (A_j) = {10, 11, 12, 13, 14}. We consider every individual has exactly attribute STR allele values. If the value of an attribute A_j is missing, then we denote the attribute value of A_j by a category $\in$ which means empty. Let X_i be individual, represented as [$X_{i,1}$, … ,$X_{i,m}$]. We define X_i = $X_{i,j}$ if $X_i = X_{k,\,j}$ for $1 \leq j \leq m$, where the relation X_i = X_k does not mean that X_i and X_k are the same individual because there exists the two individuals have equal STR allele values in attributes A_1, …. , A_m. In Y-STR, there exist a lot cases; individuals share the same STR allele values throughout markers but different individuals.

RESULTS

Experimental assembly: The experiments are led on two datasets of Y-STR data that were obtained from a database, called worldfamilies.net (www.worldfamilies.net):

- The first data set is Y-STR data for haplogroup applications.
- The second data set is Y-STR data for Y-Surname applications.

Both data sets are based on 25 markers (attributes). The data sets are as follows:

- The first data set of Y-STR haplogroup consists of 535 records. The original data were 3419 that consisted of 29 groups. See the complete data in Family Tree DNA (www.familytreedna.com). However, the data had been filtered to chose only 8 groups, called haplogroups, which consist of B (47), D (32), E (12), F (162), H (63), I (123), J (35) and N(61) respectively. The values in the parenthesis indicate the number of records belong to the particular group.

- The second data set of Y-STR Surname consists of 112 data that belong to Donald Surname.

See the details in Donald Surname Project (http://dna-project.clan-donald-usa.org) However, the original of 896 data of Donald Surname had been filtered to obtain only 112 individual based on its modal haplotypes. The modal haplotype for this surname is: 13, 25, 15, 11, 11, 14, 12, 12, 10, 14, 11, 31, 16, 8, 10, 11, 11, 23, 14, 20, 31, 12, 15, 15, 16. Thus, there are 6 classes based on the genetic distance described as mismatches 0-5. The mismatches are determined and compared between the individual and its modal haplotypes.

For better results, each dataset and algorithm is runs about 100 times. For each run, the dataset is randomly reordered from the original order. For hard *k*-Modes, the diverse method is used for initial *k* because the methods had been proved better than the distinct method (Huang, 1998).

Performances:

In general case: An external quality measure is the F measure (Aggarwal *et al*., 1999) a measure that combines the precision and recall ideas from information retrieval (Van Rijsbergen, 1989; Kowalski, 1997). We treat each cluster as if it were the result of a query and each class as if it were the desired set of documents for a query. We then calculate the recall and precision of that cluster for each given class. More specifically, for cluster j and class i:

$$\mathrm{Re}\,call(i,j) = \frac{n_{ij}}{n_i} \tag{4}$$

$$\mathrm{Pr}\,ecision(i,j) = \frac{n_{ij}}{n_j} \tag{5}$$

where,

n_{ij} = The number of members of class *i* in cluster *j*

n_j = The number of members of cluster *j*

n_i = The number of members of class *i*

The F-measure of cluster *j* and class *i* is given by:

$$F(i,j) = \frac{2*\left(\mathrm{Re}\,call(i,j)*\mathrm{Pr}\,ecision(i,j)\right)}{\mathrm{Pr}\,ecision(i,j)+\mathrm{Re}\,call(i,j)} \tag{6}$$

For an entire hierarchical clustering the F-measure of any class is the maximum value it attains at any node in the tree and an overall value for the F measure is computed by taking the weighted average of all values for the F measure as given by the following:

$$F = \sum_i \frac{n_i}{n} \max\left\{F(i,j)\right\} \tag{7}$$

where, the max is taken over all clusters at all levels and n is the number of documents

In our particular case: In order to evaluate the clustering accuracy, the misclassification matrix proposed by Huang (1998) is used to analyze the correspondence between clusters and the haplogroups or surname of the instances. Clustering accuracy is defined by:

$$clustering\,accuracy = \frac{\sum_{i=1}^{k} a_i}{n} \tag{8}$$

where,

k = The number of clusters

a_i = The number of instances occurring in both cluster *i* and its corresponding haplogroup or surname

n = The number of instances in the data sets

For precision and recall, the calculation s based on the following equations:

$$\mathrm{Pr}\,ecision = \frac{\sum_{l=1}^{k}\left(\frac{a_l}{a_l+b_l}\right)}{n} \tag{9}$$

$$\mathrm{Re}\,call = \frac{\sum_{l=1}^{k}\left(\frac{a_l}{a_l+c_l}\right)}{n} \tag{10}$$

a_I = The number of correctly classified objects

a_I = The number of incorrectly classified objects

c_I = The number of objects in a given class but not in a class

n = The number of classes/clusters

Table 1 gives overview clustering results of the evaluated algorithms. The bold faced numbers refer to the best clustering result obtained by that particular algorithm. For Y-STR 535 dataset, the highest average clustering accuracy belongs to *k*-Modes algorithm. The algorithm obtained the average of clustering accuracy, 80.38% as compared to the other algorithms: *k*-Medoids (78.19%). However, in contrast the *k*-Medoids algorithm produces a value that closes to zero for standard deviation. The algorithm also obtained the highest value of minimum accuracy of 100 runs, whereas the *k*-Modes algorithm recorded the highest value of 94.77% for maximum value of 100 runs.

For Y-STR 112 data set, the average clustering accuracy obtained by all algorithms is in between 38%-44% only. This is because all algorithms cannot work well with the objects having very strong similarity among the classes. In fact, some of the Y-STR objects

Table 1: The summary result for 100 runs of four algorithms

		Clustering algorithms	
	Evaluation (accuracy)	k-Modes	k-Medoids
Data set	Average	0.8038	0.7819
535 Y-STR	Standard deviation	0.0922	0.0262
	Max	0.9477	0.8336
	Min	0.5925	0.7514
	Average	0.4212	0.4363
	Standard deviation	0.0265	0.0149
112 Y-STR	Max	0.4643	0.4554
	Min	0.3393	0.3482

Table 2: The summary result for precision

		Clustering algorithms	
	Evaluation (accuracy)	k-Modes	k-Medoids
Data set	Average	0.7338	0.6982
535 Y-STR	Standard deviation	0.0890	0.0575
	Max	0.9000	0.7839
	Min	0.5387	0.5444
	Average	0.3857	04196
	Standard deviation	0.1064	0.0351
112 Y-STR	Max	0.6641	0.4889
	Min	0.1934	0.2010

Table 3: The summary result for recall

		Clustering algorithms	
	Evaluation (accuracy)	k-Modes	k-Medoids
Data set	Average	0.7445	0.6949
	Standard deviation	0.0905	0.0480
535 Y-STR	Max	0.8825	0.8569
	Min	0.5202	0.9988
	Average	0.3332	0.4826
	Standard deviation	0.0792	0.0484
112 Y-STR	Max	0.4889	0.6032
	Min	0.2027	0.1764

are absolutely similar throughout 25 attributes (markers). However, the representative object-based technique produced the highest value of 43.63% but for the maximum value. Overall results can be seen; the two clustering algorithms seem to be no significant difference as it merely differs about 2%-5% only.

Table 2 and 3 give some insight values of precision and recall respectively for each algorithm. The precision and recall that are very close to 1 indicate the best matching. The *K*-004Dodes algorithm initially dominates precision values, whereas the *K*-Medoids algorithm dictates the recall values.

CONCLUSION

Overall results can be concluded that K-mode algorithm is better than K-medoid in partitioning Y-STR data; In addition, K-medoid causes high time consuming and its average clustering accuracy is also less than the *K*-Modes algorithm. If the overall results of K-medoid showed that the average clustering accuracy was obviously better than the K-mode's, it could be tested for the other extended *K*-Medoids algorithms such as CLARA and CLARANS. These two algorithms are used for large data set and improved the time efficiency.

However, from the results, it shows the *K*-Modes algorithm should be chosen for further improvement. Furthermore, from the observation of Y-STR data, the patterns are made up of many occurrences, in which they can be treated as modes. In addition, the modal haplotypes that are used to measure the genetic distance is also based on the modes. However, the modal haplotypes are not necessarily modes for all cases in any given data set because the modal haplotypes are the established references by SNP methods for a group that shares a common ancestor.

In conclusion, the ideal case if the modal haplotypes can be used as the centroids, then the *k*-Modes algorithm could be improved in partitioning Y-STR data.

REFERENCES

Ahmad, A. and L. Dey, 2007. A k-mean clustering algorithm for mixed numeric and categorical data'. Data Knowl. Eng., 63: 503-527.

Aggarwal, C.C., C.S. Gates and P.S. Yu, 1999. On the merits of building categorization systems by supervised clustering. Proceedings of the 5th Conference on ACM Special Interest Group on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, California, USA, pp: 352-356.

Chaturvedi, A., P. Green and J. Carroll, 2001. K-modes clustering. J. Classification, 18: 35-55.

Chu, S.C., J.F. Roddick and J.S. Pan, 2002. An efficient k-medoids-based algorithm using previous medoid index, triangular inequality elimination criteria and partial distance search. Proceeding of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK), London, UK, pp: 63-72.

Ester, M., H.P. Kriegel and X. Xu, 1995. Knowledge discovery in large spatial databases: focusing techniques for efficient class identification. Proceeding of the International Symposium on Advances in Spatial Databases, Portland, ME, 951: 67-82.

Fitzpatrick, C., 2005. Forensic Genealogy. Rice Book Press, Fountain Valley, CA.

Fitzpatrick, C. and A. Yeiser, 2005. DNA and Genealogy. Rice Book Press, Fountain Valley, CA.

Gan, G., Z. Yang and J. Wu, 2005. A genetic *k*-modes algorithm for clustering categorical data. Lect. Notes Artif. Intell., 3584(2005): 195-202.

Gowda, K.C. and E. Diday, 1991. Symbolic clustering using a new dissimilarity measure. Pattern Recogn. Lett., 24(6): 567-578.

Hartigan, J. and M. Wong, 1979. Algorithm as136: A k-means clustering algorithm. Appl. Stat., 28: 100-108.

Haung, J.Z., M.K. Ng, H. Rong and Z. Li, 2005. Automated variable weighting in k-mean type clustering. IEEE T. PAMI, 27(5).

He, Z., S. Deng and X. Xu, 2005. Improving k-modes algorithm considering frequencies of attribute values in mode. Lect. Notes Artif. Intell., 3801(2005): 157-162.

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

Huang, Z., 2003. A note on k-modes clustering. J. Classification, 20: 257-26.

Huang, Z. and M.K. Ng, 1999. A fuzzy k-modes algorithm for clustering categorical data. IEEE T. Fuzzy Syst., 7(4): 446-452.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surveys, 31: 264-323, DOI: 10.1145/331499.331504.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. Prentice Hall Inc., Englewood Cliffs, New Jersey, pp: 320.

Kaufman, L. and P.J. Rousseeuw, 2005. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, NY.

Kaufmann, L. and P.J. Rousseeuw, 1990. Finding Group in Data: An Introduction to Cluster Analysis. John Willey and Sons, NY.

Kim, D.W., K.H. Lee and D. Lee, 2004. Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recogn. Lett., 25(11): 1263-1271.

Kim, D.W., K.Y. Lee, D. Lee and K.H. Lee, 2005. A *k* populations algorithm for clustering categorical data. Pattern Recogn., 38(7): 1131-1134.

Kowalski, G., 1997. Information Retrieval Systems: Theory and Implementation. 3rd Edn., Kluwer Academic Publishers, USA, pp: 296.

Krishna, K. and M. Murty, 1999. Genetic k-means algorithm'. IEEE T. Syst. Man Cy., 29(3): 433-439.

Ng, M.K. and J.C. Wong, 2002. Clustering categorical data sets using tabu search techniques. Pattern Recogn., 35(12): 2783-2790.

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

San, O.M., V.N. Huynh and Y. Nakamori, 2004. An alternative extension of the k-means algorithm for clustering categorical data. Int. J. Appl. Math. Comput. Sci., 14(2): 241-247.

Sun, Y., Q. Zhu and Z. Chen, 2002. An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recogn. Lett., 23(7): 875-884.

Van Rijsbergen, C.J., 1989. Information Retrieval. 2nd Edn., Buttersworth Publishers, London, UK, pp: 323.

Zhang, Q. and I. Couloigner, 2005. A new and efficient k-medoid algorithm for spatial clustering. Proceeding of the International Conference on Computational Science and Its Applications, Singapore, 3482 of LNCS: 181-189.