Centre-Based Hard and Soft Clustering Approaches for Y-STR Data

Ali Seman[a], Zainab Abu Bakar[a], Azizian Mohd. Sapawi[a]

Abstract

This paper presents Centre-based clustering approaches for clustering Y-STR data. The main goal is to investigate and observe the performance of the fundamental clustering approaches when partitioning Y-STR data. Two fundamental Centre-based hard clustering approaches, *k*-Means and *k*-Modes algorithms, and two fundamental Centre-based soft clustering approaches, fuzzy *k*-Means and fuzzy *k*-Modes algorithms were chosen for evaluation of Y-STR haplogroup and Y-STR Surname datasets. The results show that the soft *k*-Means clustering algorithm produces the best average of the clustering accuracy (99.62%) for Y-STR haplogroup data as well Y-STR surname data (97.61%). The overall results show that the soft clustering approach is better (92.11%) than the hard clustering approach (81.20%) in clustering Y-STR data. However, the approach for clustering Y-STR data should be further investigated to find the best way of achieving 100% of the clustering results.

Introduction

Centre-based clustering approaches have been found to be very efficient for clustering large and high-dimensional data sets compared to other types of clustering algorithms (Gan et al., 2007, p. 161). The main concern of the approaches is to find an appropriate centre for best dealing with convex shapes data points. *k*-Means clustering has represented a milestone for these approaches since it was first formalized by Macqueen (1967). Since then, many clustering algorithms derived and extended from the *k*-Means algorithm have been developed. For example, the *k*-Modes algorithm introduced by Huang (1998) uses a *k*-Means paradigm in order to overcome a problem of handling categorical data. *k*-Medoids (Kaufman and Rousseeuw, 1987) or Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 1990) also uses the same paradigm but it utilizes the objects themselves as prototype centroids.

Conversely, in the case of Y-STR data, specifically for Y-STR data from Y-Surname projects or in treatment of Y haplogroups, there have been no attempts to apply these clustering methods. Even, in the related bioinformatics area, dealing with DNA sequences in particular, clustering algorithms (not limited to centre-based clustering only) have made a significant contribution such as in gene-clustering such as in Eisen, et al., (1998), Moreau, et al., (2002), and Wu, et al., (2004), and tissue-clustering such as in Scherf, et al., (2000), Alizadeh et al., (2004) etc. Further, in Autosamal STR (atSTR), graph theoretic clustering has been applied in natural population study (See Bayer and May (2003)) and forensic sciences (See Cowell and Mostad(2002)).

The primary goal of this paper is to investigate the accuracy of the results of centre-based hard and soft clustering approaches when partitioning Y-STR data. The experiments focus only on the original or the fundamental hard and soft clustering algorithms in order to investigate and observe initial clustering results for Y-STR data. It is important to evaluate the fundamental approaches to clustering Y-STR data since there is presently no benchmark for comparison. Thus, this experiment will be restricted to; (1) the fundamental hard clustering of *k*-Means by Macqueen (1967) and *k*-Modes by Huang (1998) and; (2) the fundamental soft clustering of (fuzzy) *k*-Means initiallly proposed by Ruspini (1969) and formalized by Bezdek (1980) and (fuzzy) *k*-Modes by Huang and Ng (1999). Note that another hard *k*-Modes algorithm has been introduced by Chaturvedi (2001), but this algorithm has been shown to be equivalent to that of Huang (Huang, 2003). Thus, several new extensions of the *k*-Means approaches are excluded from our investigation, including Continuous *k*-Means algorithm (Faber, 1994), Compare-means algorithm (Philips, 2002), fuzzy covariance clustering (Gustafson and Kessel, 1979), Fuzzy c-Elliptotypes algorithm, (Bezdek, 1981) etc and *k*-Modes algorithm such as *k*-Modes with new dissimilarity measures by He et al., (2007) and Ng et al., (2007), *k*-Population (Kim et al., 2007), and a new fuzzy *k*-Modes proposed by Ng and Li (2009).

Address for correspondence: aliseman@tmsk.uitm.edu.my

[a]Centre for Computer Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Malaysia

Received: Mar. 26, 2010; Accepted: Aug. 17, 2010; Published: Dec. 19, 2010.

Centre Based Hard and Soft Clustering Approaches

The centre-based clustering has been evolving significantly, even though the results have not reached 100% of the clustering accuracy for all benchmarks datasets yet. The trend now has shifted from the hard clustering approaches to the soft clustering approaches. The soft clustering approaches seem to be a promising approach particularly in dealing with categorical data (See Ng and Li (2009) and Kim et al., (2007)). The hard clustering is sometimes called non-fuzzy clustering, whereas the soft clustering is referred to fuzzy clustering. From a general perspective, the hard and soft clustering can be seen as differing in the assigning of values for a partition matrix. The hard clustering approach only assigns a value of 1 or 0. In contrast, the soft clustering is more relaxed, allowing the values to be part of more than one cluster. The higher the value is, the higher the degree of confidence that the objects belong to that cluster.

The Centre-based clustering can be described as follows:

Let us suppose that the objective is to partition a data set, D into cluster, C. Suppose that k is known as a priori. Let X ={X_1, X_2,..., X_n} be set of data with set of attributes A ={A_1,A_2,..., A_m}. The partition of D, whether hard or soft partition is to minimize the cost function as Equation (1), and subject to Equation (2), (3) and (4).

$$F(W,Z) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\, d(z_l, x_i) \qquad (1)$$

Subject to:

$$\sum_{l=1}^{k} w_{li} = 1, \;\; 1 \le i \le n \qquad (2)$$

$$w_{li} \in \{0,1\}, 1 \le i \le n, 1 \le l \le k \qquad (3)$$

and:

$$0 < \sum_{i=1}^{n} w_{li} < n, \;\; 1 \le l \le k \qquad (4)$$

where $k(\le n)$ is a known number of clusters, W is a (k x n) partition matrix, Z is $[z_1, z_1, \dots, z_k] \in \mathbb{R}^{mk}$ and $d(z_l, x_i)$ is a dissimilarity measure between z_l and x_i.

The algorithm can be generalized as:

Step 1: *Choose an initial point $Z^{(1)} \in \mathbb{R}^{mk}$. Determine $W^{(1)}$ such that $F(W, Z^{(1)})$ is minimized. Set t=1.*

Step 2: *Determine $Z^{(t=1)}$ such that $F(W^t, Z^{(t+1)})$ is minimized. If $F(W^t, Z^{(t+1)}) = F(W^t, Z^{(t)})$ then stop; otherwise go to step 3.*

Step 3: *Determine $W^{(t+1)}$ such that $F(W^{(t+1)}, Z^{(t+1)})$ is minimized. If $F(W^{(t+1)}, Z^{(t+1)}) = F(W^{(t)}, Z^{(t+1)})$ then stop; otherwise set t=t+1 and go to step 2.*

From an optimization perspective, the main focus is to solve problem *P* as described by Bobrowski and Bezdek (1991). The problem *P* can be solved by iteratively solving the following two problems (Huang, 1998):

- Problem P_1: Fix $Z = \hat{Z}$ and solve the reduced problem $P(W, \hat{Z})$
- Problem P_2: Fix W = $\widehat{W}$ and solve the reduced problem $P(\widehat{W}, Z)$.

Thus, the differences between the hard clustering and the soft clustering are as follows:

- In the hard clustering, the problem P_1 is minimized by Equation (5).

$$w_{li} = \begin{cases} 1, & \text{if } d(x_i, z_l) = \min_{1 \le l \le k} d(x_i, z_l), \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

  whereas, in the soft clustering, the problem P_1 is minimized by Equation (6).

$$\widehat{w}_{li} = \begin{cases} 1, & if\ X_i = \hat{Z}_l \\ 0, & if\ X_i = \hat{Z}_h, h \ne l \\ 1 \Big/ \sum_{h=1}^{k} \left[\frac{d(\hat{Z}_l,\ X_i)}{d(\hat{Z}_h,\ X_i)}\right]^{1/(\propto -1)}, & if\ X_i \ne \hat{Z}_l \\ & and\ X_i \ne \hat{Z}_h, 1 \le h \le k \end{cases} \qquad (6)$$

- However, in the problem P_2, the hard clustering is minimized according to the *k*-Means and *k*-Mode respectively. The *k*-Means minimize $\hat{Z}$ as in Equation (7).

$$z_{lj} = \frac{\sum_{i=1}^{n} w_{li}\, x_{ij}}{\sum_{i=1}^{n} w_{li}} \qquad (7)$$

  whereas, in the *k*-Modes minimize $\hat{Z}$ as in Equation (8).

$$z_{lj} = a_j^{(r)} \qquad (8)$$

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

$$f\left(a_j^{(r)}|C_l\right) \geq f\left(a_j^{(t)}|C_l\right) \forall l, 1 \leq l \leq p_j,$$

$$a_j^{(r)} \neq a_j^{(t)} \quad (9)$$

- Further, in the soft clustering, for the Fuzzy k -Means, the minimizer $\hat{Z}$ is given by Equation (10).

$$z_{lj} = \frac{\sum_{i=1}^{n} w_{li}^{\propto} x_{ij}}{\sum_{i=1}^{n} w_{li}^{\propto}} \quad (10)$$

where $\propto \in [1, \infty)$is a weighting exponent.

In the Fuzzy k -Modes, the minimizer $\hat{Z}$ is given by Equation (11).

$$z_{lj} = a_j^{(r)} \in DOM\left(A_j\right), \text{ where}$$

$$\sum_{i,x_{i,j}=a_j^{(r)}} w_{li}^{\propto} \geq \sum_{i,x_{i,j}=a_j^{(t)}} w_{li}^{\propto} \forall l, 1 \leq t \leq n_j,$$

$$\text{for } 1 \leq j \leq m. \quad (11)$$

The main difference between the *k*-Means and *k*-Modes algorithms is that, the *k*-Means handles numerical data, whereas the *k*-Mode handles categorical data. Thus, mean is a mechanism for *k*-Means algorithm to update its centroids and mode for *k* -Modes algorithm. Consequently, the *k*-Means uses Euclidean distance as in Equation (12) and the *k*-Modes uses a simple dissimilarity measure, introduced by Kaufman and Rousseeuw (1990) as in Equation (13) and (14).

$$d_{\text{euc}}(X,Y) = \left[\sum_{j=1}^{d}\left(x_j - y_j\right)^2\right]^{\frac{1}{2}} \quad (12)$$

$$d_{sim}(X,Y) = \sum_{j=1}^{n} \delta\left(x_j, y_j\right) \quad (13)$$

where:

$$\delta\left(x_j, y_j\right) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j \end{cases} \quad (14)$$

Y-STR data

Y-STR means Short Tandem Repeats on the Y-Chromosome. The Y-STR data represents the number of times an STR motif repeats, and is called the allele value for the marker. If a Y-STR marker, say DYS391, the tandem repeats are: [TCTA] [TCTA] [TCTA] [TCTA] [TCTA] [TCTA] [TCTA], then the allele value is counted as eight. Generally, Y-STR data are used in identifying a similar group of: (1) Y-Surname and (2) Haplogroups.

Thus, the distance or the difference for persons from each other based on the Y-STR data refers to the difference in allele values for each marker. If a person shares the same allele values for each marker he is considered to be descended from the same ancestor from a genealogical perspective. A statistical method, called time-to-most-recent-common-ancestor (TMRCA) is used to evaluate how far back in time an individual shares a common ancestor. Recently, the more realistic TMRCA calculations have been proposed by Nordtvedt (2008). In a broader perspective, for instance in studying human migration patterns, it can be applied to whole haplogroups, which includes different geographical 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 typically based on 0 or 1 or 2 or 3 mismatches, whereas the haplogroups are determined by a method known as SNP analysis. All males in the world can be placed into a system of Y-DNA haplogroups named by the letters A through to T, with further subdivisions using numbers and lower case letters (www.isogg.org, 2010). The haplogroups have been established by Y-Chromosome Consortium (http://ycc.biosci.arizona.edu, 2010)

The STR data are simply sets of numerical values, so automatically the data can be treated as numerical objects. On the other hand, we observe that the occurrences of the attribute values of the STR data are more dominant. Therefore, the attribute values of STR have a tendency of mode. Thus, the STR data can also be treated as categorical data. In fact, the distance between two Y-STR objects could not be clearly measured by any common numerical distances such as Euclidean distance. For example, the mismatch distance is calculated by comparing the established modal haplotype values and the particular person STR values. However, in this clustering technique, the problem is that the modal haplotype is not part of clustering parameters.

Y-STR data as numerical and categorical values

Let $X = \{X_1, X_2, \ldots, X_n\}$ be set of *n* Y-STR data and $A = \{A_1, A_2, \ldots, A_m\}$ be set of markers/attributes of Y-STR. We define A_j is the *j*th attribute 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 they are 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 by DOM(A_j). A domain DOM(A_j) is defined as categorical if it is finite and unordered, e.g., for any missing, then we denote the attribute value of A_j by a category ϵ which means empty. Let X_i be individual, represented as $[x_{i,1}, x_{i,2},...,x_{i,m}]$. We define $X_i = X_k$, if $x_{i,j} = x_{k,j}$ 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 exist the two individuals have equal STR allele values in attributes $A_1,A_2,...,A_m$. In Y-STR, there exist a lot of cases; individuals share the same STR allele values throughout markers but different individuals.

Experimental Setup

The experiments were conducted on 2 datasets of Y-STR data. The Y-STR projects are mostly listed in a portal, named as worldfamilies.net (http://www.worldfamilies.net, 2010). The first dataset is Y-STR data for haplogroup application. The second dataset is Y-STR data for Y-Surname application. However, the datasets had been filtered to standardize on similar 25 attributes (25 markers). Further, for the surname, the datasets were filtered to obtain just the members of the main group of the family by comparing their allele values to the modal haplotype. Therefore, the final data of the surname dataset consist of the group of 0 to 5 mismatches only. Take note that the percentage of the filtered data is quite small number because: (1) the original data did not follow certain standards such as the number of markers and; (2) the data themselves were just contributed by any participant that attempts to test his/her family surname. All datasets were retrieved from the respective websites on April 2010.

The first data set consists of 267 records of Y-STR haplogroup obtained from The Finland DNA Project (http://homepage.eircom.net/~ihdp/ihdp/index.htm, 2010). The original data were 906 that consisted of 7 groups. However, the final data consist only 4 groups, called haplogroup, which are L (92), J (6), N (141), and R (28) respectively. The values in the parenthesis indicate the number of records belong to the particular group. Take note that the final data consist only the groups that have been confirmed by SNP testing only.

The second dataset consists of 236 records of four Surnames: The Donald Surname (112), The Flannery Surname (64), The Mumma Surname (42) and The William Surname (18). The description for each dataset as follows:

(i) The Donald Surname consists of 112 records obtained from the Clan Donald's DNA Projects (http://dna-project.clan-donald-usa.org, 2010). The original data were 896 records. 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.

(ii) The Flannery Surname consists of 64 records obtained from the Flannery Clan Y-DNA project (http://www.flanneryclan.ie/, 2010). The original data were 896. The modal haplotype for this surname is: 13, 24, 14, 10, 11, 14, 12, 12, 12, 14, 13, 30, 16, 9, 10, 11, 11, 26, 16, 19, 29, 15, 15, 17, 17.

(iii) The Mumma Surname consists of 42 records obtained from the Mumma-Moomaw Project (http://www.mumma.org/, 2010). The original data were 78 records. The modal haplotype for this surname is: 13, 25, 14, 11, 11, 14, 12, 12, 13, 13, 13, 29, 17, 9, 10, 11, 11, 24, 15, 19, 30, 14, 17, 17, 17.

(iv) The William Surname consists of 18 records, obtained from The Williams DNA Project (http://williams.genealogy.fm/, 2010). The original data were approximately 626 records from 94 groups. However, the final data were taken from Group 9 only. Take note that the other groups were not consistent to be considered for the final dataset. The modal haplotype for this surname is: 13, 25, 14, 11, 11, 13, 12, 12, 12, 13, 14, 29, 17, 9, 10, 11, 11, 25, 15, 18, 30, 15, 16, 16, 17.

For better results, each dataset and algorithm is run about 100 times. For each run, the dataset is randomly reordered from the original order. Further, for hard *k*-Means, the distinct initial centroids is chosen to avoid empty clustering, whereas, for hard *k*-Modes, the diverse method is used for initial *k* because the methods had been proved better than the distinct method (see Huang, 1998). In the soft clustering, the fuzziness parameter setting was set and tested from 1.1 until 2.0. For each parameter, it is tested for 100 times. However, only the parameter that produces the best clustering results was used for further analysis.

Experimental Results

This section discusses clustering results for each hard clustering; *k*-Means and *k*-Modes algorithms and each soft clustering; fuzzy *k*-Means and fuzzy *k*-Modes algorithms. Hence, this section is presented the experimental results for each algorithm: (1) clustering accuracy; (2) precision; (3) and recall.

a,b ϵ 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 the same set of attribute STR allele values. If the value of an attribute A_j is

Further, for each clustering accuracy, precision and recall, the detailed results of average, minimum, maximum and standard deviation are given. Finally, the overall performances based on three categories above are discussed. 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 as in Equation (15).

$$Accuracy = \frac{\sum_{i=1}^{k} a_i}{n} \quad (15)$$

where k, is the number of clusters, a_i is the number corresponding haplogroup or surname and n is the number of instances in the datasets.

For precision and recall, the calculation is based on Equation (16) and (17) respectively.

$$Precision = \frac{\sum_{l=1}^{k}\left(\frac{a_l}{a_l+b_l}\right)}{n} \quad (16)$$

$$Recall = \frac{\sum_{l=1}^{k}\left(\frac{a_l}{a_l+c_l}\right)}{n} \quad (17)$$

where a_l is the number of correctly classified objects; b_l is the number of incorrectly classified objects; c_l is the number of objects in a given class but not in a cluster; n is the number of classes/clusters.

Take note that for fuzzy *k*-Means and fuzzy *k*-Modes algorithms, the fuzzy parameters that produced the best clustering results were varied. For fuzzy *k*-Means algorithm, the parameter was 1.1 for the first dataset. However, in the second dataset, the parameter was 1.7. Differently, the parameter for fuzzy *k*-Modes algorithm was 1.2 for the first dataset, and the second dataset was 1.4.

Table 1 gives an overview of the clustering results of the algorithms. The bold faced numbers refer to the best clustering result obtained by the particular algorithm. Overall results show that the fuzzy *k*-Means algorithm produces the best average of the clustering accuracy for both datasets. In fact, the fuzzy *k*-Means algorithm obtained nearly 100% of the average clustering accuracy. However, during the 100-run of the experiments, the algorithm is found to be stuck at a local minimum problem while clustering the first dataset. Therefore, the algorithm obtained the same values for the minimum, the maximum and the average clustering accuracy. Unlike the first dataset, the fuzzy *k*-Means algorithm managed to cluster the second dataset without having a local minimum problem. Thus, the algorithm obtained the highest of the average clustering accuracy (97.61%) and recorded the maximum and minimum values and of 99.58% and 73.73% respectively.

In Table I, the results also show that the hard clustering approaches produce a promising clustering result. For examples, the hard *k*-Means and the hard *k*-Modes algorithms had accomplished 100% of the maximum values of the clustering accuracy for the second dataset. Furthermore, the algorithms also recorded among the highest maximum values of 99.63% and 98.88% for the first dataset. Take note that the fuzzy *k*-Modes algorithm also obtained 100% of the maximum value of the clustering accuracy. The algorithm also recorded 86.86% of the minimum clustering accuracy. These results indicate that those algorithms, especially the fuzzy *k*-Modes algorithm have a chance for further improvement. However, the fuzzy *k*-Means algorithm did not obtain the maximum value of 100%.

Tables 2 and 3 give some insight values of precision and recall respectively for each dataset. The precision and recall that are very close to 1 indicate the best matching for each pair of cluster and the corresponding haplogroup or surname classes. However, the results of the precision and recall depend much on the result of clustering accuracy. For example, if the algorithm obtained 100% of the clustering accuracy, the value of the precision and recall is automatically one. Therefore, if the algorithm produced the best clustering accuracy, it would normally obtain the best clustering results of the precision as well as the recall. As a consequence, the best results for precision and recall also belong to the fuzzy *k*-Means algorithm. However, the algorithm did not obtain the maximum value, which 1 for precision and recall for both datasets as compared to the hard *k*-Means, the hard *k*-Modes and the fuzzy *k*-Modes algorithms. For further verification, see Table II and III.

Table 4 describes on the overall performance for both approaches. The overall result has clearly shown that the soft clustering approach produces a better clustering result for both datasets. The soft clustering approach obtained about 92.11% of the overall clustering accuracy, 0.87.58 of the overall precision and 0.8666 of the overall recall as compared to its counterpart, the hard clustering approach. See Table 4 for the details.

Table 5 gives details the clustering performance, regarding the number of runs that obtained 100% of the clustering accuracy for each algorithm. Take note that each algorithm was run about 100 times. It is obviously shown that the fuzzy *k*-Modes

algorithms produced the highest number of obtaining 100% of the clustering accuracy, which are 78 times as compared to the other algorithms. Therefore, the soft clustering approach is clearly better than the hard clustering approach.

Dataset	**Evaluation (accuracy)**	**Hard Clustering**		**Soft Clustering**	
		k-Mean	*k*-Modes	*k*-Means	*k*-Modes
267 Y-STR	Average	0.7934	0.7930	**0.9962**	0.7375
	Standard Deviation	0.0825	0.0290	**-**	0.0697
	Max	**0.9963**	0.9888	0.9662	0.8165
	Min	0.6217	0.5917	**0.9662**	0.5805
236 Y-STR	Average	0.8211	0.8406	**0.9761**	0.9744
	Standard Deviation	0.1341	0.1126	**0.0311**	0.0500
	Max	**1.000**	**1.0000**	0.9958	**1.0000**
	Min	0.5381	0.5550	0.7373	**0.8686**

Table 1: The summary result for clustering accuracy of four algorithms

Dataset	**Evaluation (Precision)**	**Hard Clustering**		**Soft Clustering**	
		k-Mean	*k*-Modes	*k*-Means	*k*-Modes
267 Y-STR	Average	0.6884	0.6946	**0.9914**	0.6317
	Standard Deviation	0.1112	0.0925	**0.0000**	0.0888
	Max	**0.9914**	0.9441	**0.9914**	0.7511
	Min	0.4405	0.5793	**0.9914**	0.4894
236 Y-STR	Average	0.6620	0.6391	**0.9426**	0.9376
	Standard Deviation	0.2003	0.1879	**0.0530**	0.1217
	Max	**1.0000**	**1.0000**	0.9868	**1.0000**
	Min	0.3790	0.3790	**0.6951**	0.6902

Table 2: The summary result for precision

Dataset	**Evaluation (Recall)**	**Hard Clustering**		**Soft Clustering**	
		k-Mean	*k*-Modes	*k*-Means	*k*-Modes
267 Y-STR	Average	0.6590	0.6743	**0.9583**	0.6008
	Standard Deviation	0.1085	0.0570	**0.0000**	0.0649
	Max	0.9583	0.9548	**0.9583**	0.7863
	Min	0.4375	0.4958	**0.9583**	0.4476
236 Y-STR	Average	0.6875	0.6866	**0.9721**	0.9351
	Standard Deviation	0.1861	0.1615	**0.0554**	0.1263
	Max	**1.0000**	**1.0000**	0.9961	**1.0000**
	Min	0.3906	0.3996	0.6518	**0.6726**

Table 3: The summary result for recall

Clustering Approaches	Average		
	Accuracy	**Precision**	**Recall**
Hard Clustering	0.8120	0.6710	0.6769
Soft Clustering	**0.9211**	**0.8758**	**0.8666**

Table 4: The overall performance between the hard and soft clustering

Hard Clustering Approach		Soft Clustering Approach	
***k*-Means**	***k*-Modes**	**Fuzzy *k*-Means**	**Fuzzy *k*-Modes**
16	11	0	**78**

Table 5: The number of runs that obtained 100% of the clustering accuracy for each algorithm.

Conclusion

In this initial investigation, overall result shows that the soft clustering approach is better than the hard clustering approach in clustering Y-STR data. Therefore, the soft clustering approach can be used and proposed for further improvement in clustering Y-STR data. However, the clustering method of the fuzzy *k*-Means algorithm is found to be problematic for certain Y-STR dataset. The algorithm falls into a local minimum problem while clustering Y-STR data. This is due to: (1) the mean method as the represented centroids may not be appropriate for certain Y-STR data (2) the use of Euclidean distance as the dissimilarity measure may not also be suitable for Y-STR data. Alternatively, the fuzzy *k*-Modes algorithm can be considered for future improvement in clustering Y-STR data. The reasons are: (1) the algorithm has obtained a promising result of the overall clustering performance, especially for the second dataset. In fact, the *k*-Modes-type algorithms, regardless the hard or the soft approaches have produced the best of the correctly clustered Y-STR data as compared to the *k*-Means-type algorithms (See Table 5 for comparison). Furthermore, the fuzzy *k*-Means algorithm did not obtain even once of the correctly clustered Y-STR data for both datasets. (2) The mode method as the represented centroids for the fuzzy *k*-Modes algorithm could be used to estimate the appropriate modal haplotype for any arbitrary group of Y-STR data (3) and finally, the simple dissimilarity measure used by *k*-Modes-type algorithms is principally a similar technique when comparing Y-STR data and their modal haplotype to determine the genetic mismatches. Thus, a further investigation should be carried out to figure out those issues. Nevertheless, the results presented here can serve as a benchmark for future works in clustering Y-STR data. This attempt would open a new era for Y-STR data as the method has been introduced and used in the evaluation of haplogroups and Surname applications.

Acknowledgement

This research is part of our main research of the DNA kinship analyses funded by Fundamental Grant Research Scheme (FRGS), Ministry of Higher Education of Malaysia (Ref. no. 600-IRDC/ST/FRGS.5/3/1293; Project Code: 211201070005). Firstly, we thank Research Management Institute (RMI), Universiti Teknologi MARA (UiTM) Malaysia for their full support of this research. Secondly, we would like to extend our gratitude to many contributors toward the completion of this paper especially the dedication of our research assistants; Mr. Zahari, Miss Hasmarina and Mr. Syahrul. Finally, we would like to thank also the editor, Dr. Bettinger and the reviewer for their constructive comments and efforts to recommend this paper for publication.

References

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403: 503-511.

Bayer J, May B (2003) A graph-theoretic approach to the partition of individuals into full-sib families, Molecular Ecology. 12: 2243-2250

Bezdek J.C (1980) A convergence theorem for fuzzy ISODATA clustering algorithm. IEEE Transaction on Pattern Analysis and Machine Intelligence. 2: 1-8.

Bezdek, J (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press

Bobrowski L, Bezdek JC (1991) c-Means clustering with the l_1 and l^{∞} norms. IEEE Transactions on Systems, Man and Cybernetics. 21(3): 545-554.

Chaturvedi A, Foods K, Green JE (2001) K-modes Clustering. Journal of Classification, 18: 35–55.

Cowell RG, Mostad P (2002) A clustering algorithm using DNA marker information for sub-pedigree reconstruction. Journal of Forensics Sciences. 48: 1239-1248.

Eisen MB, Spellman PT, Brown PO, Botsein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of National Academy Sciences of United States of America, (Genetic). 95: 14863-14868.

Faber V (1994) Clustering and the continuous *k*-means algorithm. Los Alamos Science, 22: 138-144

Gan G, Ma C, Wu J (2007) Data clustering: Theory, algorithms, and applications. Society for Industrial and Applied Mathematics (SIAM): Philadelphia, VA.

Gustafson DE, Kessel, WC (1979). Fuzzy clustering with a Fuzzy Covariance Matrix. Proceedings IEEE on Decision and Control. 761–766

He Z, Xu X, Deng S (2007) Attribute Value Weighting in k-Modes Clustering, Computer Science e-Prints: arXiv:cs/0701013v1 [cs.AI], Cornell University Library, Cornell University, Ithaca, NY, USA, http://arxiv.org/abs/cs/0701013v1, 1-15.

http://homepage.eircom.net/~ihdp/ihdp/index.htm

http://dna-project.clan-donald-usa.org

http://www.flanneryclan.ie

http://www.mumma.org/

http://williams.genealogy.fm

http://www.isogg.org/tree/ISOGG_YDNATreeTrunk08.html

http://www.worldfamilies.net

http://ycc.biosci.arizona.edu/

Huang Z (1998) Extensions to the *k*-Means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2: 283–304.

Huang Z (2003) A Note on K-modes Clustering. Journal of Classification, 20: 257–261

Huang Z, Ng M (1999) A Fuzzy *k*-Modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems. 7:446-452.

Kaufman L and Rousseeuw PJ (1987) Clustering by means of medoids. Elsevier, 405-416.

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons: New York.

Kim DW, Lee KY, Lee D, Lee KH (2005) *k*-populations algorithm for clustering categorical data. Pattern Recognition, 38: 1131–1134.

Macqueen JB (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 281–297.

Moreau Y, De Smet F, Thijs G, Marchal K, Moor BD (2002) Functional bioinformatics of microarray data: from expression to regulation. Proceeding of IEEE, 90: 1722-1743

Ng MK, Jing L (2009) A new fuzzy *k*-modes clustering algorithm for categorical data. International Journal of Granular, Rough Sets and Intelligent Systems. 1: 105-119.

Ng MK, Li MJ, Huang JZ, He Z, (2007) On the impact of dissimilarity measure in k-modes clustering algorithm, IEEE Transactions of Pattern Analysis and Machine Intelligence. 29: 503-507.

Nordtvedt K (2008) More realistic TMRCA calculations. Journal of Genetic Genealogy, 4: 96-103.

Philips S (2002) Acceleration of *k*-means and related clustering algorithms. In Mount, D and Stein, C, editors, ALENEX: International workshop on algorithm engineering and experimentation, LNCS, 2409: 166-177. San Francisco: Springer-Verlag.

Ruspini ER (1969) A new approach to clustering. Information and Control. 19: 22-32.

Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN (2000) A gene expression database for the molecular pharmacology of cancer, Nature Genetic. 24: 236-244.

Wu CJ, Fu Y, Murali TM, Kasif S (2004) Gene expression module discovery using gibbs sampling. Genome Informatics, 5: 239-248.