Research Journal of Information Technology

ISSN 1815-7432

Academic Journals Inc.

www.academicjournals.com

Research Journal of Information Technology 7 (2): 112-120, 2015
ISSN 1815-7432 / DOI: 10.3923/rjit.2015.112.120

Performance Evaluations of κ-Approximate Modal Haplotype Type Algorithms for Clustering Categorical Data

[1,2]Ali Seman, [2]Azizian Mohd Sapawi and [1]Mohd Zaki Salleh
[1]Integrative Pharmacogenomics Institute (iPROMISE), Level 7, FF3, Universiti Teknologi MARA (UiTM), Puncak Alam Campus, 42300, Bandar Puncak Alam, Selangor, Malaysia
[2]Center for Computer Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Selangor, Malaysia

Corresponding Author:Ali Seman, Integrative Pharmacogenomics Institute (iPROMISE), Level 7, FF3, Universiti Teknologi MARA (UiTM), Puncak Alam Campus, 42300, Bandar Puncak Alam Selangor, Malaysia Tel: +60355211191 Fax: +60355435100

ABSTRACT

The effectiveness of the performance of κ-Approximate Modal Haplotype (κ-AMH)-type algorithms for clustering Y-short tandem repeats (Y-STR) of categorical data has been demonstrated previously. However, newly introduced κ-AMH-type algorithms, including the new κ-AMH I (Nκ-AMH 1), the new κ-AMH II (Nκ-AMH II) and the new κ-AMH III (Nκ-AMH III), are derived from the same κ-AMH optimization and fuzzy procedures but with the inclusion of two new methods, namely, new initial center selection and new dominant weighting methods. This study evaluates and presents the performance of κ-AMH-type algorithms for clustering five categorical data sets-namely, soybean, zoo, hepatitis, voting and breast. The performance criteria include accuracy, precision and recall analyses. Overall, κ-AMH-type algorithms perform well when clustering all of the categorical data sets mentioned above. Specifically, the N κ-AMH I algorithm exhibits the best performance when clustering the five categorical data sets; this algorithm obtained the highest combined mean accuracy score (at 0.9130), compared to those of κ-AMH (0.8971), N κ-AMH II (0.8885) and N κ-AMH III (0.9011). This high score is associated with the newly introduced initial center selection, combined with the original dominant weighting method. These results present a new and significant benchmark, indicating that κ-AMH-type algorithms can be generalized for any categorical data.

Key words: Fuzzy clustering algorithms, categorical data, partitioning methods, optimization

INTRODUCTION

Clustering algorithms have been used for categorical data since, Huang (1998) introduced the hard κ-Modes algorithm. A year later, an algorithm called the fuzzy κ-Modes algorithm (Huang and Ng, 1999) based on fuzzy clustering procedures was introduced. These algorithms use the same optimization procedure as the κ-means algorithm. However, they replace the mean with the mode as the center of clusters and they replace the Euclidean distance with a simple dissimilarity measure. Consequently, the κ-Mode-type algorithm became a pillar algorithm for clustering categorical data. Many extended κ-Mode-type algorithms were later developed e.g., κ-Modes with four new weighting attributes (He *et al*., 2007), κ-Modes with a new dissimilarity measure (Ng *et al*., 2007), κ-Population (Kim *et al*., 2005) and the new fuzzy κ-Modes algorithm (Ng and Jing, 2009).

Recently, an algorithm called κ-Approximate Modal Haplotype (κ-AMH) was introduced exclusively for clustering Y-Short Tandem Repeat (Y-STR) categorical data (Seman *et al*., 2012a). This algorithm is also based on the fuzzy clustering procedure; however, it is quite different. The main difference is the use of objects (also known as medoids) as the center of clusters for the κ-AMH algorithm instead of the mode mechanism (known as centroid), imposed by the κ-Modes-type algorithms. Besides that, the algorithm uses the difference optimization technique, which is the maximization of its cost function. As a result, the medoid mechanism of κ-AMH algorithm improved the overall accuracy scores for clustering Y-STR categorical data; particularly the minimum and maximum accuracy scores improved compared to its competitor, the κ-Modes-type algorithms. The detailed comparisons are presented by Seman *et al.* (2012a).

Furthermore, the κ-AMH algorithm was also tested for clustering other categorical data sets-e.g., soybean, voting, breast, zoo, mushroom, lymphography, credit, hepatitis and adults with promising results (Seman *et al*., 2013b). In fact, in some cases (e.g., for clustering the zoo data set with a higher number of clusters (seven clusters)), the κ-AMH algorithm was more accurate than κ-Mode-type algorithms such as the fuzzy κ-Modes and the new fuzzy κ-Modes algorithm. From both the results discussed above, it is clear that the main advantage of the κ-AMH algorithm is that it is not overly sensitive to the initial center selections, even though they are randomly selected. It was already known that the initial center selection is one of the factors contributing to the performance of clustering results, particularly for κ-means-type algorithms (Li *et al*., 2008).

Moreover, in a recent attempt to improve the clustering results of Y-STR data, three new κ-AMH algorithms were developed as extended κ-AMH algorithms. These include the new κ-AMH I (Nκ-AMH I), the new κ-AMH II (Nκ-AMH II) and the new κ-AMH III (Nκ-AMH III) (Seman *et al*., 2015). These κ-AMH-type algorithms are derived from the κ-AMH algorithm by maintaining the same clustering procedure with two newly introduced methods: (1) New initial center selection method and (2) New dominant weighting method. As a result, the Nκ-AMH III algorithm demonstrated a 2% improvement to the clustering accuracy when clustering Y-STR categorical data, as compared to that of the κ-AMH algorithm and the other two κ-AMH-type algorithms (viz., Nκ-AMH I and Nκ-AMH II). This improvement is, in fact, contributed by the two methods above. Detailed results for κ-AMH-type algorithms can be found in Seman *et al.* (2015).

This study aims to evaluate the performance of the newly introduced κ-AMH-type algorithms (Nκ-AMH I, II and III) when clustering five categorical data sets namely; soybean, zoo, hepatitis, voting and breast comparing the respective performance of these new algorithms with the original κ-AMH algorithm. These new κ-AMH algorithms and in particular Nκ-AMH III, have demonstrated superiority when clustering Y-STR data. However, the characteristics of Y-STR categorical data are essentially different from common categorical data such as soybean, zoo, hepatitis, voting and breast. The main difference is that the degree of similarity among common categorical data is not high as it is with Y-STR data. Common categorical data sets typically comprise both similar and distinct objects but not identical objects. By contrast, Y-STR data sets comprise identical, similar and distinct objects, because each identical object refers to a different individual with the same DNA characteristics. Further details regarding Y-STR data are available in Seman *et al.* (2013a). It is important to know that the new κ-AMH algorithms with the two new methods designed specifically for Y-STR categorical can also be applied efficiently for the common categorical data listed above. Thus, the favorable results obtained in this study present a new and significant benchmark for clustering categorical data.

MATERIALS AND METHODS

κ-AMH algorithm: The original κ-AMH algorithm begins to randomly find k clusters in X categorical objects as the initial center selection H (the medoid). For each medoid, h∈H is tested for each object x∈X one-by-one. The final center cluster H is obtained by replacing h by x if the cost function P(W, D) is maximized, as in Eq. 1:

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

P(W, D) is calculated in Eq. 2 as:

$$P(W,D) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{\alpha} d_{li} \tag{2}$$

Where:

- $w_{li}^{\alpha} \in W$ is a (k×n) fuzzy membership matrix that denotes the degree of fuzzy membership for the object i in the l^{th} cluster, which contains a value between 0 and 1, as in Eq. 3:

$$w_{li}^{\alpha} = \begin{cases} 1 & X_i = H_l \\ 0 & X_i = H_z, z \neq l \\ \left[\sum_{z=1}^{k} \frac{d(X_i, H_i)}{d(X_i, H_z)}\right]^{\frac{-\alpha}{\alpha-1}} & \text{Otherwise} \end{cases} \tag{3}$$

where, k (≤n) is a known number of clusters, H is the medoid, α∈[1, ∞) is a weighting exponent and $d(X_i, H_z)$ is the distance measured between the object X_i and the medoid H_z:

- $d_{li} \in D$ is another (k×n) partition matrix with a dominant weighting value of 1.0 or 0.5. The dominant weighting value d_{li} is calculated as follows:

$$d_{li} = \begin{cases} 1.0, & \text{if } w_{li}^{\alpha} = \max^{w_{li}^{\alpha}, 1 \leq l \leq k} \\ 0.5, & \text{otherwise} \end{cases} \tag{4}$$

Subject to:

$$1.5 \leq \sum_{l=1}^{k} d_{li} \leq k, 1 \leq i \leq n \tag{5}$$

$$0.5 < \sum_{i=1}^{n} d_{li} < n, 1 \leq l \leq k \tag{6}$$

The descriptions above represent the simplified version of the κ-AMH algorithm reported by Seman *et al.* (2015). However, the original descriptions of the κ-AMH algorithm that are similar to those described above can be found in Seman *et al.* (2012a).

κ-AMH-type algorithms: Nκ-AMH I, II and III are the new κ-AMH algorithms that maintain the original κ-AMH optimization procedure but replace two methods:

- New initial center selection method: This method replaces the randomized initial center by selecting any identical or similar center selections. The detailed steps and descriptions of this method can be found in Seman *et al.* (2015)
- New dominant weighting method: This method replaces the original dominant weighting method for the κ-AMH algorithm, based on a 50 or 100% probability with a weighting method based on three probabilities, 50, 75 and 100%. An additional value of 0.75 (75%) was proposed to appropriately handle any maximum membership values ranging from 1/κ to 1.0, where κ is the number of clusters. The new dominant weighting value method is described in Eq. 7

$$d_{li} = \begin{cases} 1.00 \text{ if } w_{li}^{\alpha} = \max^{w_{li}^{\alpha}\,1\le l\le k} = 1.00 \\ 0.75 \text{ if } w_{li}^{\alpha} = \max^{w_{li}^{\alpha}\,1\le l\le k} > 1/k \\ 0.50 \quad \text{otherwise} \end{cases} \tag{7}$$

The detailed steps and descriptions for this method can be found in Seman *et al.* (2015). For clarity, Table 1 shows the main differences between κ-AMH-type algorithms.

Benchmark data sets: For bench marking the results, five categorical data sets were used to compare the clustering performances of the algorithms. They include the soybean, zoo, hepatitis, voting and breast data sets from the UCI repository (Lichman, 2013). Table 2 provides a summary of all the data sets and presents the number of objects, classes and attributes for each set.

Evaluation methods: The analyses were primarily based on mean accuracy scores obtained from experiments that were run 100 times (100 run) for each algorithm and data set. Each data set was

Table 1: κ-AMH-type algorithms

κ-AMH-type algorithm	Initial center selection method	Dominant weighting method	Note
Nκ-AMH I	New method	Old method	Combining the new center selection method with the original κ-AMH's dominant weighting method
Nκ-AMH II	Old method (randomized)	New method	Combining the original center selection (randomized selection) method with the new dominant weighting method
Nκ-AMH III	New method	New method	Combining the new center selection method with the new dominant weighting method

Table 2: Summary of categorical data sets

Data set	No. of objects	No. of classes	No. of attributes
Soybean	47	4	21
Zoo	101	7	17
Hepatitis	155	2	13
Voting	435	2	16
Breast	699	2	9

randomly re-ordered before each run. In order to calculate the accuracy scores, a misclassification matrix was used. The accuracy was calculated and defined by Huang (1998) as:

$$Ac = \frac{\sum_{l=1}^{\kappa} a_l}{n} \tag{8}$$

where, κ is the number of clusters, a_l is the number of instances occurring in cluster l and its corresponding group and n is the number of instances in the data sets. As a secondary analysis, the precision and recall scores were calculated as:

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

$$Rc = \frac{\sum_{l=1}^{\kappa}\left(\frac{a_l}{a_l + c_l}\right)}{n} \tag{10}$$

where, b_l is the number of incorrectly classified objects and c_l is the number of objects in a given class but not in a cluster. The accuracy, precision and recall scores close to 1 indicate the best matching for each cluster and corresponding class pair.

RESULTS

Table 3 lists the clustering accuracy for each data set. In general, clustering categorical data using these κ-AMH-type algorithms produced very promising results. Overall, the score differences among the κ-AMH-type algorithms were merely in the range 1-2%. However, the Nκ-AMH I algorithm obtained the highest combined mean accuracy score at 0.9130 which is slightly higher

Table 3: Mean, minimum and maximum clustering accuracy scores for all data sets after 100 experimental runs performed on each data set

	Data set					
Accuracy and algorithm	1	2	3	4	5	Combined mean
Mean						
κ-AMH	0.9762	0.9421	0.7898	0.8752	0.9023	0.8971
Nκ-AMH I	1.0000	0.9604	0.7871	0.8620	0.9557	0.9130
Nκ-AMH II	0.9981	0.8834	0.7831	0.8618	0.9160	0.8885
Nκ-AMH III	1.0000	0.9168	0.7871	0.8603	0.9413	0.9011
Minimum						
κ-AMH	0.9362	0.8713	0.6375	0.8552	0.8526	0.8306
Nκ-AMH I	1.0000	0.9604	0.7871	0.8598	0.9557	0.9126
Nκ-AMH II	0.9575	0.6931	0.7742	0.8575	0.7668	0.8098
Nκ-AMH III	1.0000	0.8812	0.7871	0.8575	0.9413	0.8934
Maximum						
κ-AMH	1.0000	1.0000	0.8625	0.8897	0.9471	0.9399
Nκ-AMH I	1.0000	0.9604	0.7871	0.8667	0.9557	0.9140
Nκ-AMH II	1.0000	0.9703	0.7871	0.8713	0.9413	0.9140
Nκ-AMH III	1.0000	0.9604	0.7871	0.8667	0.9413	0.9111

than κ-AMH (0.8971), Nκ-AMH II (0.8885) and Nκ-AMH III (0.9011). This excellent performance was essentially demonstrated by its consistency, whereby, from the 100-runs, the Nκ-AMH I obtained identical accuracy scores for data set 1, 2, 3 and 5 (Table 3). This was because of the incorporation of the new initial center selection method (a fixed selection) into the Nκ-AMH algorithm. From the experimental results, it can be seen that the new initial center selection method, when combined with the original dominant weighting method, resulted in significantly improved clustering results. In fact, this combination significantly increased the highest minimum combined mean accuracy score (0.9126).

Table 4 lists the results from a secondary analysis, using precision and recall scores. This analysis was used to provide insight into the clustering results. From the precision analysis, the Nκ-AMH I algorithm produced the highest mean precision scores for all data sets except data set 3. This indicates that the algorithm was also more precise than the other new κ-AMH type algorithms, because approximately 85.71% of the objects were correctly clustered but with a slightly lower precision score than the κ-AMH algorithm (0.8639). Thus, it yielded only 14.29% of objects that were not supposed to be in the clusters-the second-best score after κ-AMH (13.61%) and considerably better than Nκ-AMH II (17.59%) and Nκ-AMH III (16.51%). By contrast, in the recall analysis, the Nκ-AMH I algorithm had a slightly higher recall score (0.8712) than the κ-AMH algorithm (0.8674). This means that Nκ-AMH I yielded the lowest score, with only 12.88% of the objects that were not supposed to be in clusters, compared to κ-AMH (13.26%), Nκ-AMH II (16.37%) and Nκ-AMH III (15.21%).

Table 5 shows the results of a group comparison of clustering accuracy scores between Nκ-AMH I and the other three κ-AMH-type algorithms using one-way ANOVA analysis. The assumption of homogeneity of variance was violated (Levene $F(3, 1996) = 20.14$, $p < 0.05$); therefore, the Welch F-ratio was reported. There was a significant variance in clustering accuracy scores among the four algorithms ($F(3, 1107.42) = 8.382$, $p<0.05$); therefore, the Games Howell procedure was used for a multiple group comparison. The performance of Nκ-AMH I (M = 0.9130, 95% CI [0.9062, 0.9199]) differed significantly from the κ-AMH and Nκ-AMH II ($p<0.05$) algorithms. However, the results from the Nκ-AMH I algorithm did not differ significantly from those of Nκ-AMH III ($p = 0.0647$).

Table 4: Clustering precision and recall scores for all data sets and the combined data set

	Data set					
Algorithm	1	2	3	4	5	Combined mean
Mean (precision)						
κ-AMH	0.9766	0.8992	0.6522	0.8695	0.9221	0.8639
Nκ-AMH I	1.0000	0.9286	0.5306	0.8751	0.9514	0.8571
Nκ-AMH II	0.9981	0.8406	0.5209	0.8734	0.8873	0.8241
Nκ-AMH III	1.0000	0.8423	0.5306	0.8737	0.9277	0.8349
Mean (recall)						
κ-AMH	0.9831	0.9105	0.6916	0.8866	0.8654	0.8674
Nκ-AMH I	1.0000	0.9325	0.6163	0.8568	0.9505	0.8712
Nκ-AMH II	0.9984	0.8183	0.5792	0.8559	0.9295	0.8363
Nκ-AMH III	1.0000	0.8263	0.6163	0.8554	0.9416	0.8479

Table 5: One-way ANOVA comparison between N κ-AMH I and the other three κ-AMH-type algorithms

Accuracy				
(I) Algorithm	(J) Algorithm	Mean difference (I-J)	S.E	p-value
Nκ-AMH I	κ-AMH	0.0159	<0.0047	0.0039
	Nκ-AMH II	0.0246	<0.0050	<0.0001
	Nκ-AMH III	0.0119	<0.0048	0.0647

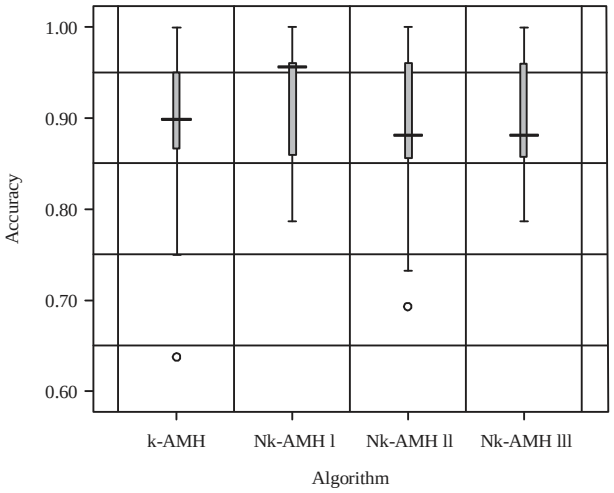

Fig. 1: Box plot comparison of the clustering accuracy performance of κ-AMH-type algorithms. The Nκ-AMH I algorithm shows excellent performance, with the highest median and minimum accuracy scores

Statistically, the Nκ-AMH I algorithm outperformed both κ-AMH and Nκ-AMH II and its performance was comparable to that of Nκ-AMH III.

In addition, Fig. 1 shows a visual representation of the accuracy scores for the κ-AMH-type algorithms. The Nκ-AMH I algorithm was clearly superior, with the highest median and minimum accuracy scores in comparison with the other κ-AMH-type algorithms. The highest median and minimum accuracy scores were effectively backed by the identical accuracy scores obtained for data sets 1, 2, 3 and 5. However, the Nκ-AMH I and Nκ-AMH III algorithms appeared to have similar box plot diagram representations, which indicate similar mean performances as proven by the ANOVA analysis above, although the median value was relatively low. In fact, the diagram shows that both algorithms (Nκ-AMH I and III) are more consistent than the other two algorithms (κ-AMH and Nκ-AMH II), without producing any outliers and they resulted in the highest values in terms of their minimum accuracy scores. This indicates that the new initial center selection significantly contributes to the overall performances of both algorithms.

DISCUSSION

The overall performance indicates that κ-AMH-type algorithms, including the original κ-AMH (Seman *et al*., 2012a) and the extended κ-AMH *viz*., Nκ-AMH I, II and III (Seman *et al*., 2015), can cluster any categorical data. For the common categorical data used in the evaluation above, new κ-AMH algorithms such as Nκ-AMH I and III demonstrated an improvement of 1-2% in terms of the clustering accuracy when compared with the original κ-AMH algorithm. This indicates that the improved methods in the new κ-AMH algorithms significantly contributed to the overall improvement. In general, the new initial center selection method and the new dominant weighting method introduced by Seman *et al*. (2015) were not restricted by highly similar categorical data (such as Y-STR data (Seman *et al*., 2013a) that had previously been applied and reported by Seman *et al*. (2010a, b, c, d, e, f, 2012a, b, 2015). Meanwhile, the applications of these methods can also be extended for use with common categorical data and reported with promising results, particularly for a high number of clusters such as the seven clusters of zoo data set (Seman *et al*., 2013b). In addition, for the new κ-AMH-type algorithm represented by Nκ-AMH I, which combines the new initial center selection method with the original dominant weighting

method, achieved a clustering accuracy score that was 2% higher than the other algorithms, as evinced by the experiment above. In contrast, when clustering Y-STR data, the Nκ-AMH III combining the new initial center selection method with the new weighting dominant method achieved the highest clustering accuracy score overall (Seman *et al*., 2015). This new initial center selection is, in reality, a fixed initial selection method. Thus, it differs fundamentally from the original procedure of seeding the centroids with a randomized initial selection method. Moreover, from the experimental results, it can be concluded that the original dominant weighting method with two conditions (100 and 50%) incorporated in Nκ-AMH 1 is the most suitable method for common categorical data. However, the new dominant weighting method with three conditions (100, 75 and 50%) incorporated in Nκ-AMH III and combined with the fixed initial center selection is the most suitable method for clustering Y-STR categorical data. This indicates that the new initial center selection method (i.e., the fixed initial center selection method) plays an important role in contributing to the performance of clustering categorical data, regardless of the type of data (whether Y-STR or common categorical data).

The previous results reported for Y-STR data, coupled with the experimental results presented above, confirm the algorithms' generality when clustering categorical data. In other words, the algorithms are not exclusively limited to single applications using Y-STR data; rather, they can be used for any application of categorical data. The κ-AMH-type algorithms are a type of approximation algorithm that require an optimal solution, therefore, an improvement of approximately 1-2%, as obtained by Nκ-AMH I and III, is considerably good. Therefore, both Nκ-AMH I and Nκ-AMH III can be generalized and used for clustering any categorical data. Notably, all κ-AMH-type algorithms can be used to further develop clustering tools. This is because κ-AMH-type algorithms, characterized by the use of the object as the center cluster and highlighted in Seman *et al*. (2013b), have their own advantages over κ-Mode-type algorithms (e.g., the κ-mode (Huang, 1998), fuzzy κ-mode algorithms (Huang and Ng, 1999) and new fuzzy κ-mode algorithms (Ng and Jing, 2009), which are characterized by the mode mechanism.

CONCLUSION

From the analyses of the experimental results, it can be seen that the κ-AMH variants significantly contributed to the process of clustering categorical data. The combination of the newly introduced methods incorporated in the κ-AMH procedures demonstrated their advantages over certain categorical data. Although, the combination of the two methods represented by the Nκ-AMH III algorithm did not result in a consistently superior performance for all data sets, as it did with the Y-STR data sets, this merely indicates that each method had its own advantages. Thus, the κ-AMH algorithm and its variants offer unique advantages for clustering categorical data. These algorithms have the potential to be generalized that is, they can be used to cluster Y-STR data and any categorical data, such as the data presented above.

ACKNOWLEDGMENT

This study was supported by the Fundamental Research Grant Scheme, Ministry of Education, Malaysia. We would like to thank IRMI and UiTM for their support with this research. We also extend our gratitude to those who have contributed toward the completion of this study.

REFERENCES

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

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

Kim, D.W., K.Y. Lee, D. Lee and K.H. Lee, 2005. A κ-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. IEEE Trans. Knowledge Data Eng., 20: 1519-1534.

Lichman, M., 2013. UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA.

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

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

Seman, A., Z.A. Bakar and A. M. Sapawi, 2010a. Modeling centre-based hard and soft clustering for Y chromosome short tandem Repeats (YSTR) data. Proceedings of the International Conference on Science and Social Research, December 5-7, 2010, Kuala Lumpur, Malaysia, pp: 68-73.

Seman, A., Z.A. Bakar and A.M. Sapawi, 2010b. Attribute value weighting in k-modes clustering for Y-Short Tandem Repeats (Y-STR) surname. Proceedings of the International Symposium on Information Technology, June 15-17, 2010, Kuala Lumpur, Malaysia, pp: 1531-1536.

Seman, A., Z.A. Bakar and A.M. Sapawi, 2010c. Centre-based clustering for Y-Short Tandem Repeats (Y-STR) as numerical and categorical data. Proceedings of the International Conference on Information Retrieval and Knowledge Management, March 17-18, 2010, Shah Alam, Selangor, pp: 28-33.

Seman, A., Z.A. Bakar and A.M. Sapawi, 2010d. Centre-based hard and soft clustering approaches for Y-STR data. J. Genet. Geneal., 6: 1-9.

Seman, A., Z.A. Bakar and A.M. Sapawi, 2010e. Centre-based hard clustering algorithm for Y-STR data. Malaysian J. Comput., 1: 62-73.

Seman, A., Z.A. Bakar and A.M. Sapawi, 2010f. Hard and soft updating centroids for clustering Y-Short Tandem Repeats (Y-STR) data. Proceedings of the IEEE Conference on Open Systems, December, 5-7, 2010, Kuala Lumpur, Malaysia, pp: 6-11.

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

Seman, A., Z.A. Bakar and M.N. Isa, 2012b. Evaluation of κ-Modes-Type Algorithms for Clustering Y-Short Tandem Repeats Data. Trends Bioinf., 5: 47-52.

Seman, A., Z.A. Bakar and M.N. Isa, 2013a. First Y-short tandem repeat categorical dataset for clustering applications. Dataset Papers Sci., 10.7167/2013/364725

Seman, A., Z.A. Bakar, A.M. Sapawi and I.R. Othman, 2013b. A Medoid-based method for clustering categorical data. J. Artif. Intell., 6: 257-265.

Seman, A., A.M. Sapawi and M.Z. Salleh, 2015. Towards development of clustering applications for large-scale comparative genotyping and kinship analysis using Y-short tandem repeats. J. Integr. Biol., 19: 361-367.