Trends in Bioinformatics

ISSN 1994-7941

science
alert

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

Trends in Bioinformatics 5 (2): 47-52, 2012
ISSN 1994-7941 / DOI: 10.3923/tb.2012.47.52

Evaluation of *k*-modes-type Algorithms for Clustering Y-Short Tandem Repeats Data

[1]Ali Seman, [1]Zainab Abu Bakar and [2]Mohamed Nizam Isa
[1]Center for Computer Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia
[2]Faculty of Medical, Masterskill University College of Health Sciences, No. 6, Jalan Lembah, Bandar Seri Alam, 81750 Johor Bahru, Johor, 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

ABSTRACT

This paper reports on the experimental results of the *k*-modes-type algorithms for partitioning Y-Short Tandem Repeats (Y-STR) data. The results were based on the clustering accuracy scores of five hard and three soft *k*-modes-type algorithms. Six Y-Short Tandem Repeats data sets were used as a benchmark for the evaluation. The results clearly indicated that the soft *k*-modes-type clustering algorithms are the most reliable algorithms for partitioning Y-STR data.

Key words: Clustering, data mining, categorical algorithms, Y-STR data

INTRODUCTION

Y-Short Tandem Repeats (Y-STR) is generally useful for distinguishing lineages and providing information about lineage relationships (Kayser *et al*., 2004). For examples, the Y-STR is used to trace similar groups of Y-Surname projects as to support the traditional genealogical study e.g., (Perego *et al*., 2005; Perego, 2005; Hutchison *et al*., 2004). Furthermore, in forensic genetics, the Y-STR is one of the primary concerns in human identifications e.g., in sexual assault cases (Dekairelle and Hoste, 2001; Betz *et al*., 2001; Corach *et al*., 2001) paternity testing (Rolf *et al*., 2001; Jobling *et al*., 1997), missing person (Dettlaff-Kakol and Pawlowski, 2002), human migration pattern (Stix, 2008) and rediscovering ancient cases (Gerstenberger *et al*., 1999; Foster *et al*., 1998).

Recently, some of the *k*-modes-type algorithms have been applied for clustering Y-STR data such as *k*-modes algorithms (Seman *et al*., 2010a, b), Fuzzy *k*-modes algorithms (Seman *et al*., 2010b, c), *k*-modes algorithms with the attribute weighting schemas (Seman *et al*., 2010d) and New Fuzzy *k*-modes and *k*-population algorithms (Seman *et al*., 2010e). The applications above have previously shown promising results. Thus, this paper tends to evaluate *k*-modes-type algorithms for clustering Y-STR data to serve as a benchmark for future applications and improvements.

The pillar of the partitional algorithms is the *k*-Means clustering algorithm which has been introduced since four decade ago by MacQueen (1967). A lot of extended versions of *k*-means paradigm including the *k*-modes algorithm (Huang, 1998) for categorical data have been introduced. The introduction of the *k*-modes algorithm was due to the ineffectiveness of *k*-means algorithm to handle the categorical data. Since then, the *k*-modes algorithm has become the center of intention in solving categorical data problems.

Thus, various *k*-modes-type algorithms were introduced including hard *k*-modes-type algorithms e.g., the *k*-modes with RVF (He *et al.*, 2007; Ng *et al.*, 2007; San *et al.*, 2004), the *k*-modes with UAVM (He *et al.*, 2007), the *k*-modes with Hybrid 1 (He *et al.*, 2007), the *k*-modes with hybrid II (He *et al.*, 2007) and soft *k*-modes-type algorithms e.g., the fuzzy *k*-modes algorithm (Ng *et al.*, 2007), the *k*-population algorithm (Kim *et al.*, 2005) and the new fuzzy *k*-modes algorithm (Ng and Jing, 2009).

MATERIALS AND METHODS

Y-STR data: The Y-STR data were mostly obtained from a database called worldfamilies.net (www.worldfamilies.net). The first, second and third data sets represented the Y-STR data for haplogroup applications, whereas the fourth, fifth and sixth data sets represented the Y-STR data for Y-Surname applications. All data sets had been filtered to standardize on similar 25 attributes (25 markers). The chosen markers were: DYS393, DYS390, DYS19 (394), DYS391, DYS385a, DYS385b, DYS426, DYS388, DYS439, DYS389I, DYS392, DYS389II, DYS458, DYS459a, DYS459b, DYS455, DYS454, DYS447, DYS437, DYS448, DYS449, DYS464a, DYS464b, DYS464c and DYS464b. Furthermore, for the surname, the data sets 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 data sets consisted of the group of 0 to 5 mismatches only. All data sets were retrieved from the respective websites on April 2010. The details about the data sets are as follows:

- The first data set consists of 751 objects of Y-STR haplogroup that belongs to the Ireland Y-DNA project (www.familytreedna.com/public/IrelandHeritage/). The data consists of only 5 haplogroups, which are E (24), G (20), L (200), J (32) and R (475)
- The second data set consists of 267 objects of Y-STR haplogroup obtained from the Finland DNA Project (http://www.familytreedna.com/public/Finland/). The data consists of only 4 haplogroups, which are L (92), J (6), N (141) and R (28)
- The third data set consists of 263 objects obtained from Y-Haplogroup project (www.worldfamilies.net/yhapprojects/). The data consists of Group G (37), Group N (68) and Group T (158)
- The fourth data set consists of 236 objects that combined four Surnames: The Donald Surname (http://dna-project.clan-donald-usa.org/), The Flannery Surname (http://www.flanneryclan.ie/), The Mumma Surname (http://www.mumma.org/) and The William Surname (http://williams.genealogy.fm/)
- The fifth data set consists of 112 objects belongs to the Philips DNA project (http://www.phillipsdnaproject.com/). The data consists of 8 family groups: Group 2 (30), Group 4 (8), Group 5 (10), Group 8 (18), Group 10 (17), Group 16 (10), Group 17 (12) and Group 29 (7)
- The sixth data set consists of 112 objects belong to Brown Surname project (http://brownsociety.org). The data consists of 14 family groups: Group 2 (9), Group 10 (17), Group 15 (6), Group 18 (6), Group 20 (7), Group 23 (8), Group 26 (8), Group 28 (8), Group 34 (7), Group 44 (6), Group 35 (7), Group 46 (7), Group 49 (10) and Group 91 (6)

The values in the parenthesis indicate the number of objects belong to that particular group.

Algorithms: Five hard and three soft *k*-modes-type algorithms were used to evaluate the Y-STR data. The hard *k*-modes-type algorithms refer to the *k*-modes algorithm, the *k*-modes with RVF, the *k*-modes with UAVM, the *k*-modes with hybrid 1 and *k*-modes with hybrid II. Furthermore, the soft *k*-modes-type algorithms refer to the fuzzy *k*-modes algorithm, the *k*-population algorithm and the new fuzzy *k*-modes algorithm.

Evaluation method: The misclassification matrix proposed by Huang (1998) was used to analyze the correspondence between clusters and the haplogroups or surname of the instances. Thus, the value from the misclassification matrix was used to obtain the clustering accuracy scores in order to evaluate the clustering performances. Thus, the performance of the algorithms was measured by the clustering accuracy, r as defined by Huang (1998) as Eq. (1):

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \tag{1}$$

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

RESULTS AND DISCUSSIONS

The hard *k*-modes-type algorithms: Table 1 shows the clustering results for the hard *k*-modes-type algorithms. The boldface numbers indicate the highest clustering accuracy scores. Overall results shows that the *k*-modes-RVF obtained the highest clustering accuracy scores for the data set 1, 2, 3, 5 and 6. For the data set 4, the *k*-modes-UAVM produced the highest score of 0.87. From this result, the extended *k*-modes algorithm using relative frequency method is seen to be more promising than the others. However, none of the algorithm produced an optimal solution (1.00) for all the data sets.

The soft *k*-mode-type algorithms: Table 2 shows the clustering results for the soft *k*-modes-type algorithms. The results obviously show that the *k*-population algorithm performs better than the other algorithms. It produced the highest clustering accuracy scores for all data sets. In fact, for the

Table 1: Clustering accuracy scores for the hard *k*-modes-type algorithms

	Data set					
Algorithms	1	2	3	4	5	6
k-modes	0.70	0.79	0.84	0.84	0.74	0.62
k-modes-RVF	**0.79**	**0.83**	**0.87**	0.78	**0.87**	**0.72**
k-modes-UAVM	0.65	0.75	0.83	**0.87**	0.56	0.54
k-modes-HI	0.56	0.81	0.85	0.77	0.80	0.64
k-modes HII	0.56	0.82	0.83	0.79	0.81	0.70

Bold face number indicate the higher clustering scores

Table 2: Clustering accuracy scores for the soft *k*-modes-type algorithms

	Data set					
Algorithms	1	2	3	4	5	6
Fuzzy *k*-modes	0.56	0.74	0.74	0.97	0.76	0.66
k-Population	**0.80**	**0.90**	**0.97**	**1.00**	**0.97**	**0.84**
New fuzzy *k*-modes	0.71	0.84	0.77	**1.00**	0.77	0.69

Bold face number indicate the higher clustering scores

Table 3: Clustering accuracy scores of the hard vs. soft *k*-modes-type algorithms

				95% Confidence interval for mean			
Type	N	Mean	SD	Lower bound	Upper bound	Min	Max
Hard	3000	0.75	0.14	0.75	0.76	0.38	1.00
Soft	**1800**	**0.82**	**0.15**	**0.81**	**0.82**	**0.32**	**1.00**

Bold face number indicate the higher clustering scores

data set 4, the algorithm has achieved an optimal solution of 1.00. Furthermore, the new fuzzy *k*-modes algorithm has also produced an optimum solution. Take note that the data set 4 consists of four different surname: the Donald Surname, the Flannery Surname, the Mumma Surname and the William Surname. In this case, both algorithms managed to cluster the data perfectly.

The hard vs. Fuzzy *k*-mode-type algorithms: Table 3 shows that the soft *k*-modes-type algorithms is the most reliable approach for clustering Y-STR data. The soft-type algorithms produced 0.82 of accuracy scores as compared 0.75 for the hard-type algorithms. However, the soft-type algorithms have also produced a lower score of the minimum accuracy score (0.32) as compared to the hard-type algorithms (0.38). It was due to the randomized initial centroid selection used by the soft-type algorithms that failed to produce a good initial centroid. In addition, the Y-STR data were characterized by a lot of similar data that would lead to poor initialization centroids. This problem did not occur to the hard-type algorithms because they were using the diverse method proposed by Huang (1998) for obtaining the distinct objects for the initializations.

CONCLUSIONS

From the experiments above, the results showed that the *k*-modes-type algorithms can be used for partitioning similar groups of Y-STR data. However, there is a room for further improvement in terms of optimizing the clustering results. Both types, the hard and soft *k*-modes-type algorithms produced lower scores for the minimum accuracy scores. It was due to the characteristic of Y-STR data that contained a lot of similar and almost similar objects. This characteristic leads to the production of poor initial centroid selections. From the result above, the improvement can be done through the soft *k*-modes-type algorithm. The *k*-population algorithm or the new fuzzy *k*-modes algorithm could be the preferred algorithm for further improvement.

ACKNOWLEDGMENTS

We would like to extend our gratitude to many contributors toward the completion of this paper including Mr. Azizian Mohd Sapawi, Ida Rosmini and our research assistants; Syahrul, Azhari, Kamal, Hasmarina, Nurin, Soleha, Fadzila, Suhaida and Shukriah.

REFERENCES

Betz, A., G. Baâler, G. Dietl, X. Steil, G. Weyermann and W. Pflug, 2001. DYS STR analysis with epithelial cells in a rape case. Forensic Sci. Int., 118: 126-130.

Corach, D., L.F. Risso, M. Marino, G. Penacino and A. Sala, 2001. Routine Y-STR typing in forensic casework. Forensic Sci. Int., 118: 131-135.

Dekairelle, A.F. and B. Hoste, 2001. Application of a Y-STR-pentaplex PCR (DYS19, DYS389I and II, DYS390 and DYS393) to sexual assault cases. Forensic Sci. Int., 118: 122-125.

Dettlaff-Kakol, M.A. and R. Pawlowski, 2002. First polish DNA manhunt: An application of Y-chromosome STRs. Int. J. Legal Med., 116: 289-291.

Foster, E.A., M.A. Jobling, P.G. Taylor, P. Donnelly and P. de Knijff *et al.*, 1998. Jefferson fathered slaves last child. Nature, 396: 27-28.

Gerstenberger, J., S. Hummel, T. Schultes, B. Hack and B. Herrmann, 1999. Reconstruction of a historical genealogy by means of STR analysis and Y-haplotyping of ancient DNA. Eur. J. Hum. Genet., 7: 469-477.

He, Z., X. Xu and S. Deng, 2007. Attribute value weighting in *k*-modes clustering. Cornell University Library, Cornell University, Ithaca, NY., USA. http://arxiv.org/abs/cs/0701013

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

Hutchison, L.A.D., N.M. Myres and S.R. Woodward, 2004. Growing the family tree: The power of DNA in reconstructing family relationships. Proceedings of the 1st Symposium on Bioinformatics and Biotechnology, September 24, 2004, Colorado Springs, CO., USA., pp: 42-49.

Jobling, M.A., A. Pandaya and C. Tyler-Smith, 1997. The Y-chromosome in forensic analysis and paternity testing. Int. J. Legal Med., 110: 118-124.

Kayser, M., R. Kittler, A. Erler and M. Hedman, 2004. A comprehensive survey of human Y-chromosomal microsaatellites. Am. J. Hum. Genet., 74: 1183-1197.

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.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. Proc. Berkeley Symp. Math. Statist. Prob., 1: 281-297.

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: 105-118.

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.

Perego, U.A., 2005. The power of DNA: Discovering lost and hidden relationships in world library and information congress. Proceedings of the 71th IFLA General Conference and Council, August 14-18, 2005, Oslo Norway,.

Perego, U.A., A. Turner, J.E. Ekins and S.R. Woodward, 2005. The science of molecular genealogy. Nat. Geneal. Soc. Quart., 93: 245-259.

Rolf, B., W. Keil, B. Brinkmann, L. Roewer and R. Fimmers, 2001. Paternity testing using Y-STR haplotypes: Assigning a probability for paternity in cases of mutations. Int. J. Legal Med., 115: 12-15.

San, O.M., V. N. Huynh and Y. Nakamori, 2004. An alternative extension of the k-Means algorithm for clustering categorical data. Int. J. Applied Math. Comput. Sci., 1472: 241-247.

Seman, A., Z.A. Bakar and A. M. Sapawi, 2010a. Modeling center-based hard and soft clustering for Y chromosome short tandem Repeats (Y-STR) 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. Center-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. Center-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. 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.

Stix, G., 2008. Traces of the distant past. Sci. Am., 299: 56-63.