IJRSET JUNE 2017 Volume 4, Issue 6 Pages: 28-33
ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

NOVEL CLUSTERING METHOD FOR THE CATEGORICAL DATA USING MATHEMATICAL FUZZY PARTITIONING

[1]Dr.R.Rangaraj
[1]Head & Associate Professor ,PG & Research Dept of Computer Science ,
[1]Hindusthan College of Arts & Science , TamilNadu, India.

ABSTRACT: Past subtractive clustering techniques can be utilized for numerical information, yet it can't be connected to downright information since trait estimations of clear cut information don't have a characteristic requesting. The k-modes clustering calculation is without a doubt a standout amongst the most broadly utilized partitioned calculations for straight out information. By utilizing a basic coordinating difference measure for all out articles and modes rather than means for cluster, another approach is created, which permits the utilization of the k-implies worldview to productively group huge clear cut informational collections. A fluffy k-modes calculation is introduced and the viability of the calculation is shown with exploratory outcomes.

Keywords: [fuzzy partitioning, k-means algorithm, subtractive clustering method, k-modes.]

1. INTRODUCTION

Clustering is a standout amongst the most vital tasks in exploratory information examination. Its essential objectives are to aggregate the comparative examples into a similar group and finding the important structure of the information. Clustering has a long and rich history in an assortment of logical controls including human sciences, science, prescription, brain science, measurements, arithmetic, designing, and software engineering. Conventional parcel clustering algorithms incorporate k-means, k-modes, et cetera. K-means clustering is without a doubt the most generally utilized partitioned clustering algorithm, and its affectability of instatement of clustering has caught the consideration of the clustering groups for a significant long time. The conventional mountain clustering algorithm and this method is one exceptionally basic and extremely successful and can be utilized on evaluating the bunch focuses. The customary mountain clustering views some framework focuses as bunch focuses and these focuses have the biggest mountain work esteem. Yet, this mountain clustering method in some cases may cause the expanding computational intricacy.

Despite the fact that the conventional k-means clustering algorithm, the subtractive clustering approach or their enhanced methods can gain the clustering execution, they just can be utilized for numerical information and they are not reasonable for unmitigated information. The primary reason is that characteristic estimations of all out information don't have a characteristic requesting. To do clustering for clear cut information, some clustering methods were proposed. For instance, k-modes, which originated from the possibility of the general k-means algorithm, can execute the clustering method for the absolute information. Besides, others of the broadly utilized and exhibited some clustering methods for straight out information are as per the following: fuzzy k-modes, fuzzy k-meloids .The fundamental consequence of this paper is to give a method to locate the fuzzy bunch modes when the basic coordinating disparity

measure is utilized for downright protests. The fuzzy rendition has enhanced the k-modes algorithm by allocating certainty to objects in various groups. These certainty esteems can be utilized to choose the centre and limit objects of cluster, in this way giving more valuable data to managing limit objects.

SCCA roused by the k-modes and subtractive clustering methods connected Hamming separation to processing the mountain work an incentive rather than Euclidean separation .A few tests are executed on a few UCI genuine datasets and some exploratory outcomes exhibit this proposed approach can acquire the more fulfilled clustering exactnesses than the customary k-modes utilizing Hamming separation.

2. THE NEW SCCA ALGORITHM

The subtractive clustering method is employed to do clustering for categorical data and propose the SCCA algorithm. Assume that one dataset X , _ _ 1, , n X _ x _ x , is a data set for clustering in Rd . One procedure for the proposed SCCA algorithm is as follows:

I. Let k be the number of cluster modes and Let t _1.

II. Compute one mountain function's value for every data i x from X . Here Eq.(1) is the mountain function.

$$M_{t-1}(x_i) = \sum_{j-1}^{n}\left(-D\left(x_j, x_i\right)\right) \quad \blacktriangleleft 1$$

III. Select one data sample xt 1 from X as a cluster mode, and this data sample xt 1 as the calculated highest mountain function value.

IV. If t _ k , then SCCA stops; otherwise goto V.

V. For each data sample i x , discount the mountain function value of the using Eq.(2).

$$M_t(x_i) = M_{t-1}(x_i) - M_{t-1}\left(-D(x^{t-1}, x_i)\right) \blacktriangleleft 2$$

VI. Let t = t +1.

The distinction between the SCCA and the traditional subtractive clustering method(SCM) is that the mountain function is not computed by means of using the traditional Euclidean distance, but Hamming distance.

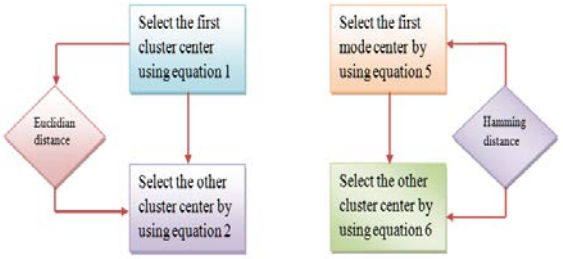

Figure 1- The distinction between SCM and SCCA

3. HARD AND FUZZY -MODES ALGORITHMS

The hard - modes algorithm, first presented in [7], has made the accompanying alterations to the - means algorithm: 1) utilizing a straightforward coordinating difference measure for downright questions; 2) supplanting the means of groups with the modes; and 3) utilizing a recurrence based method to discover the modes to tackle Problem (P2). These alterations have evacuated the numeric-just restriction of the - means algorithm yet keep up its productivity in clustering huge absolute informational collections [7].The simple matching dissimilarity measure between and is defined as follows:

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

Where

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

It is easy to verify that the function defines a metric space on the set of categorical objects. Traditionally, the simple matching approach is often used in binary variables which are converted from categorical variables. The k-modes algorithm uses the -means paradigm to cluster categorical data. The objective of clustering a set of n categorical objects into k clusters is to find w and z that minimize

$$F_e(W,Z) = \sum_{t=1}^{k}\sum_{i=1}^{n} \omega_{li}^{\alpha} d_c(Z_t, Z_i)$$

with other conditions same as in (1). Here, z represents a set of k modes for k clusters.

4. EXPERIMENTAL RESULTS

To assess the execution of this introduced SCCA, SCCA is contrasted and the mainstream algorithm k-modes on three UCI genuine datasets[21], for example, Contraceptive Method Choice(CMC), Car Evaluation(Car) and Tic-Tac-Toe Endgame (TTTE). Table I records principle qualities of these three datasets. On each dataset, Hamming separation was utilized for SCCA and k-modes, and the clustering correctness of k-modes were arrived at the midpoint of on 20 autonomous runs. one device for assessment of the clustering correctness was connected to all tests and this instrument utilizes the Hungarian method to assess the clustering exactnesses.

Datasets	The Amount of data	The amount of attributes	The Amount of Classes
TTTE	958	9	2
CMC	1473	8	3
Car	1728	6	4

TABLE I- THREE DATASETS

Datasets	K-modes	SSCA
TTTE	55.6	59.2
CMC	39.3	42.7
Car	36.4	46.6

TABLE II -THE COMPARISON OF SOME CLUSTERING ACCURACIES (%)

The clustering exactnesses are accounted for in Table II. Take note of that the best execution for each dataset is marked in intense face in Table II. The test consequences of Table II exhibit that this novel SCCA algorithm can gain the more fulfilled clustering correctness than the customary k-modes algorithm on each dataset.

The execution and productivity of the fuzzy - modes algorithm and contrast it and the reasonable means algorithm and the hard - modes algorithm, we did a few trial of these algorithms on both genuine and manufactured information. The test outcomes are talked about beneath. We utilized the three clustering algorithms to group this informational index into four cluster. The underlying means and modes were haphazardly chosen k particular records from the informational index. For the applied k-means algorithm, we initially changed over various straight out qualities into paired properties, utilizing zero for non appearance of a class and one for nearness of it. The double estimations of the qualities were then regarded as numeric esteems in the k-means algorithm.

$$r = \frac{\sum_{t=1}^{k} \alpha 1}{n}$$

A clustering result was measured by the clustering exactness r characterized as, where a1 was the quantity of cases happening in both bunch l and its relating class and was the quantity of cases in the informational collection. In our numerical tests k is equivalent to four. Every algorithm was run 100 times. Table II gives the normal exactness (i.e., the normal rates of the effectively ordered records more than 100 keeps running) of clustering by every algorithm and the normal focal preparing unit (CPU) time utilized. Fig. 1 demonstrates the disseminations of the quantity of keeps running as for the quantity of records accurately ordered by every algorithm. The general clustering execution of both hard and fuzzy k-modes algorithms was superior to anything that of the applied k-means algorithm. Also, the quantity of keeps running with right characterizations of more than 40 records (r>0.87) was significantly bigger from both hard and fuzzy - modes algorithms than that from the calculated - means algorithm. The fuzzy k-modes algorithm somewhat outflanked the hard k–modes algorithm in the general execution. The normal CPU time utilized by the k-modes-sort algorithms was significantly littler than that by the theoretical k-means algorithm. To examine the contrasts between the hard and fuzzy

k- modes algorithms, we looked at two clustering comes about delivered by them from a similar starting modes.

Table III gives the modes of four cluster delivered by the two algorithms. The modes gotten with the two algorithms are not indistinguishable. This demonstrates the hard and fuzzy k-modes algorithms without a doubt deliver distinctive cluster. By looking at the correctness of the two clustering comes about, we found that the quantity of records accurately grouped by the hard - modes algorithm was 43 while the quantity of records effectively arranged by the fuzzy - modes algorithm was 45. For this situation, there was 4.2% expansion of exactness by the fuzzy - modes algorithm. We discovered such an expansion happened by and large. Be that as it may, in a couple cases, the clustering comes about delivered by the hard k-modes algorithm were superior to those by the fuzzy –modes algorithm (see Figure-2).

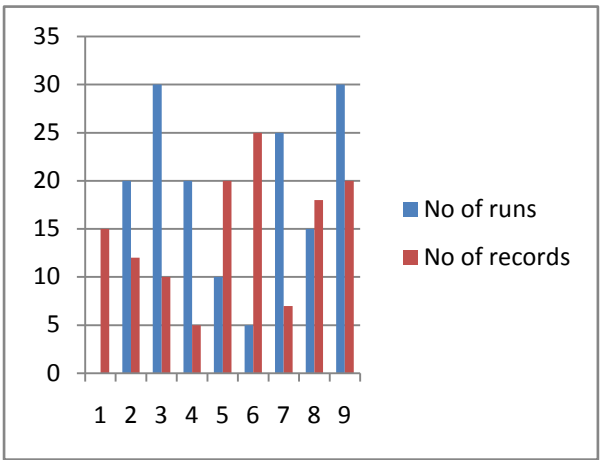

Figure 2 -The conceptual version of the k-means algorithm.

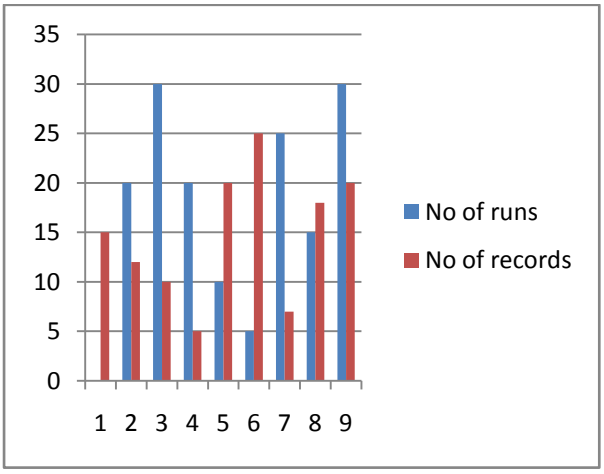

Figure 3- The hard k-modes algorithm.

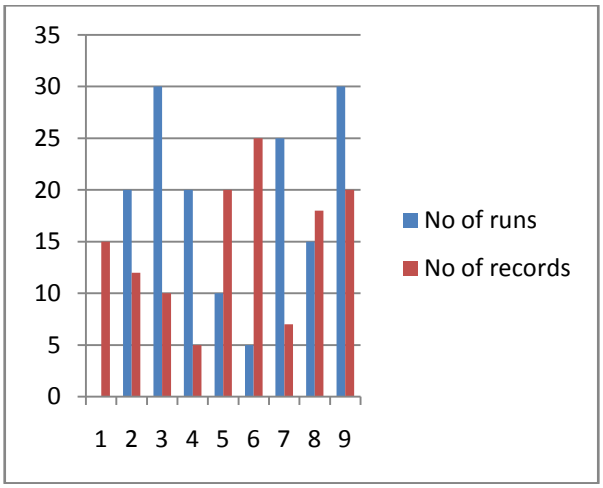

Figure 4- The fuzzy k-modes algorithm.

	Conceptual k-means	Hard k-modes	Fuzzy k-modes
Accuracy	0.74	0.78	0.79
CPU time in seconds	0.164	0.24	0.03

Table III - The average clustering and CPU time in seconds by different clustering methods.

It gives the modes of four clusters produced by the two algorithms. The modes obtained with the two algorithms are not identical. This indicates that the hard and fuzzy -modes algorithms.

CONCLUSION

Our paper portrays one Novel clustering method for the unmitigated information. This method can apply Hamming separation to computing the mountain work an incentive rather than the Euclidean separation in the conventional subtractive clustering method. In analyses, the proposed method has demonstrated its prevalence more than three datasets. We have presented the fuzzy k-modes algorithm for clustering unmitigated articles in view of expansions to the fuzzy k–means Algorithm. The test comes about have demonstrated that the - modes-sort algorithms are successful in recuperating the inalienable clustering structures from clear cut information if such structures exist. In addition, the fuzzy segment grid gives more data to help the

client to decide the last clustering and to recognize the limit objects. Such data is amazingly helpful in applications, for example, information mining in which the questionable limit articles are at times more intriguing than items which can be grouped with sureness. In my future work, a few parameters for this exhibited method will be set to make the unsupervised clustering's execution better, and utilize more informational indexes and assessing methods for this novel approach.

REFERENCES

[1]A K Jain, M N Murty, P J Flynn. Data clustering: A review [J]. ACM Computing Surveys, 999, 31(3): 264-323.

[2]S R Kannan, S Ramathilagam, P C Chung. Effective fuzzy C-means clustering algorithms for data clustering problems [J]. Expert Systems with Applications, 2012, 39: 6292-6300.

[3]A K Jain. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[4]L Jing, M K Ng, J Z Huang. An entropy weighting K-means algorithm for subspace clustering of high-dimensional sparse data [J]. IEEE Transactions on Knowledge and Data Engineering, USA: Institute of Electrical and Electronics Engineers, 2007, 19(8): 1026-1041.

[5]M Emre Celebi, Hassan A Kingravi, Patricio A Vela. A comparative study of efficient initialization methods for the K-means clustering algorithm [J]. Expert Systems with Applications, 2013, 40: 200-210.

[6] J. C. Gower, "A general coefficient of similarity and some of its properties," BioMetrics, vol. 27, pp. 857–874, 1971.

[7] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining Knowledge Discovery, vol. 2, no. 3, Sept. 1998.

[8] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[9] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data—An Introduction to Cluster Analysis. New York: Wiley, 1990.

[10] T. Kohonen, Content-Addressable Memories. Berlin, Germany: Springer-Verlag, 1980.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Symp. Mathematical Statistics and Probability, Berkeley, CA, 1967, vol. 1, no. AD 669871, pp. 281–297.

[12] R. S. Michalski and R. E. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy," IEEE Trans.Pattern Anal. Machine Intell., vol. PAMI-5, pp. 396–410, July 1983.

[13] H. Ralambondrainy, "A conceptual version of the k-means algorithm," Pattern Recognition Lett., vol. 16, pp. 1147–1157, 1995.

[14] R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in Proc. 20th Very Large Databases Conf., Santiago, Chile, Sept. 1994, pp. 144–155.

[15] E. R. Ruspini, "A new approach to clustering," Inform. Contr., vol. 19, pp. 22–32, 1969.

[16] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality," IEEE Trans. Pattern Anal. Machine Intell., vol. 6, pp. 81–87, Jan. 1984.

[17] M. A. Ismail and S. Z. Selim, "Fuzzy c-means: Optimality of solutions and effective termination of the problem," Pattern Recognition, vol. 19, no. 6, pp. 481–485, 1986.

[18] M. A. Woodbury and J. A. Clive, "Clinical pure types as a fuzzy partition," J. Cybern., vol. 4-3, pp. 111–121, 1974.

[19] H. Jia, Y.M. Cheung, J.M. Liu, "A new distance metric for unsupervised learning of categorical data," IEEE Transactions on Neural Networks and Learning Systems, Vol.27, No.5, pp.1065-1078, 2016.

[20] G. Gan, C. Ma, J. Wu, "Data clustering: theory algorithms, and application," SIAM Series on Statistics and Applied Probability, VA, 2007.

[21] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.
[22]Tool for evaluation of the clustering performance,http://www.cad.zju.edu.cn/home/dengcai/Data/Clustering.html
[23] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," IEEE Transactions on Knowledge and Data Engineering, Vol.17, pp.1624-1637, 2005.