http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 7 (2): 356-360, 2008
ISSN 1812-5638
© 2008 Asian Network for Scientific Information

A Weighted Mean Subtractive Clustering Algorithm

[1]JunYing Chen, [1,2]Zheng Qin and [1,3]Ji Jia
[1]Department of Computer Science and Technology, Xi'an JiaoTong University, Xi'an, China
[2]School of Information Science and Technology, Tsinghua University, Beijing, China
[3]Air Force Equipment Academy, Beijing, China

Abstract: In this study, we propose a weighted mean subtractive clustering algorithm in which new cluster centers are derived by using weighted mean method on the data points around the center prototypes found by subtractive clustering. Comparisons between weighted mean subtractive clustering and other clustering alogrithms are performed on three datasets by using three indexes and visual methods. The experimental results show that weighted mean subtractive clustering finds more reasonable cluster centers and groups data better than other clustering alogrithms do.

Key words: Subtractive clustering, weighted mean, cluster center

INTRODUCTION

Clustering has emerged as a popular technique in many areas, including data mining, statistics and machine learning, etc. Clustering analysis groups the data into clusters so that objects within a cluster have high similarity in comparison to another. A variety of clustering algorithms have been proposed, including k-means, fuzzy c-means (Bezdek *et al*., 1999) and mountain clustering (Yager and Filev, 1994) algorithms.

Mountain clustering estimates the cluster centers by constructing and destroying the mountain function on a grid space. However, the mountain method is computed in the amount of computation growing exponentially with the increase in the dimensionality of the data. Subtractive clustering (Chiu, 1994) was proposed to reduce the computational cost by computing the mountain function on the data points rather than the grid nodes. Nikhil and Chakraborty (2000) stated that subtractive clustering is computationally less expensive than mountain clustering. But the results may be less accurate due to selection of cluster centers only from dataset. Yang and Wu (2005) improved the subtractive clustering by modifying mountain function and revised mountain function to automatically estimate the parameters in accordance with the structure of the data and also the number of clusters and Kim *et al.* (2005) proposed a kernel-induced distance instead of the conventional distance when calculating the mountain value of data point.

In this study, we propose weighted mean subtractive clustering in which the cluster centers are derived by using weighted mean method on the data points in a certain hypercube. The point with higher potential has more influence on the cluster center than the one with lower potential. The weighting coefficient of the data point is based on the proportional of its potential to the sum of potential of the data points surrounding center prototype found by subtractive clustering. Comparative experiments between weighted mean subtractive clustering and other clustering algorithms were executed on three datasets.

SUBTRACTIVE CLUSTERING

The subtractive clustering algorithm is described as follows:

Consider a group of n data points $\{x_1,x_2,...,x_n\}$, where, x_i is a vector in the feature space. Without loss of generality, we assume that the feature space is normalized so that all data are bounded by a unit hypercube. We consider each data point as a potential cluster center and define a measure of the point to serve as a cluster center. The potential of x_i, denoted as P_i, is computed by Eq. 1.

$$P_i = \sum_{j=1}^{n} \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{(r_a/2)^2}\right) \quad (1)$$

where, r_a is a positive constant defining a neighborhood radius, $\|\ \|$ denotes the Euclidean distance. A data point with many neighboring data points will have a high potential value and the points outside r_a have little influence on its potential.

Corresponding Author: JunYing Chen, Department of Computer Science and Technology, Xi'an JiaoTong University, Xi'an, China

The first cluster center c_1 is chosen as the point having the highest potential. The potential of c_1 is denoted as PotVal(c_1). Next, the potential of each data point x_i is revised as follows:

$$P_i = P_i - PotVal(c_1)\exp(-\frac{\|x_i - c_1\|^2}{(r_b/2)^2}) \quad (2)$$

where, $r_b = 1.5r_a$ is usually set to avoid obtaining closely spaced cluster centers. The data points near the first cluster center will have greatly reduced their potential and will unlikely be selected as the next cluster center. After the potential of all data points have been reduced according to Eq. 2, the one with the highest potential is selected as the second cluster center. Then, the potential of the remaining points is again reduced. Generically, after the kth cluster center c_k is determined, the potential is revised as follows:

$$P_i = P_i - PotVal(c_k)\exp(-\frac{\|x_i - c_k\|^2}{(r_b/2)^2}) \quad (3)$$

where, c_k is the location of the kth cluster center and PotVal(c_k) is its potential value. The process continues until the stopping criterion defined in (Li *et al*., 1999) is reached.

From the clustering process, two conclusions can be drawn:

- The point with high potential has more chance to be selected as cluster center than the point with less potential. Each cluster center is a point with relatively high potential.
- Cluster centers are selected only from the data points whether or not the actual cluster centers are in the dataset. However, the actual cluster centers are not necessarily located at one of the data points.

WEIGHTED MEAN SUBTRACTIVE CLUSTERING

The weighted mean subtractive clustering algorithm consists of the following steps:

Step 1: Compute the potential of each data point using Eq. 1; set the number of cluster centers as k = 1.

Step 2: Select the point with the highest potential denoted as c_k, the data points surrounding c_k with radius smaller than r_a are denoted as $(x_1^{(k)}, x_2^{(k)}, \ldots x_{m(k)}^{(k)})$. Then, the weighted mean cluster center $\overline{c_k}$ is computed as follows:

$$\overline{c_k} = \frac{\sum_{j=1}^{m(k)} PotVal(x_j^{(k)}) * x_j^{(k)}}{\sum_{j=1}^{m(k)} PotVal(x_j^{(k)})} \quad (4)$$

where, m(k) is the number of data points surrounding c_k with radius smaller than r_a.

Step 3: The potential of each data point is revised as follows:

$$P_i^{(k+1)} = P_i^{(k)} - PotVal(\overline{c_k})\exp(\frac{\|x_i - \overline{c_k}\|^2}{(r_b/2)^2}) \quad (5)$$

Where:

$$PotVal(\overline{c_k}) = \sum_{i=1}^{n}\exp(-\frac{\|\overline{c_k} - x_i\|^2}{(r_a/2)^2}) - \sum_{j=1}^{k-1}\exp(-\frac{\|\overline{c_k} - \overline{c_j}\|^2}{(r_b/2)^2})$$

Step 4: If the stop criterion is met, then stop the process; otherwise, set k = k + 1, return to Step 2.

In weighted mean subtractive clustering, the location of the cluster center is decided by not only one data point but all data points in a neighboring area. The point with high potential has a comparatively big impact on the cluster center. A measure of the influence is based on its potential, the more potential, the more influence.

EXPERIMENTS AND DISCUSSION

Description of the dataset: Experimental studies are performed on one artificially created dataset named as ArtData and two real life datasets named as Iris and Pima available from machine learning database at UCI (Blake and Merz, 1998).

ArtData: This dataset contains two clusters in 2D. The data points of cluster 1 are random numbers uniformly generated in an area enclosed by a circle with centre (0.2, 0.2) and radius 0.2. The data points of cluster 2 are random numbers uniformly generated in an area between two circles with radius 0.1 and 0.2, respectively. Cluster 1 contains 100 samples and cluster 2 contains 200 samples. Figure 1 shows the distribution of the dataset.

Iris: The Iris dataset contains 150 samples with 4 attributes. It has three classes and each class contains 50 samples. One of the three classes is well separated from the other two, which are not easily separable due to the overlapping of their convex hulls.

Pima: Pima is a medical dataset used for identifying the diabetes disease in patients. It contains 768 samples with 8 attributes belonging to two classes.

Because ouly 2-D or 3-D clustering problems can be visually inspected, we ouly presented the distribution of cluster centers of dataset 1. Visual representation of dataset 2 and dataset 3 can not be presented because of their high number of dimensions. In Fig. 1, the centers found by Weighted Mean Subtractive Clustering (WMSC), Kernel-based Subtractive Clustering (KSC), Subtractive Clustering (SC) and Mountain Clustering (MC) were plotted to clearly show the effectiveness of WMSC to create new cluster centers. From Fig. 1, we could see that WMSC found more reasonable centers than KSC and SC did because WMSC can create new centers which are not included in original dataset. MC found grid nodes as centers which are closest to the centers found by WMSC.

To demonstrate the ability of finding centers with high potential by our proposed algorithm, we used WMSC, KSC, SC and MC to the three datasets. All four clustering algorithms work on the same principle and find cluster centers based on the potential of data points or grid nodes. KSC computes the potential by using kernel-induced distance. But the other three algorithms use conventional distance. For comparison, the potential of centers found by KSC was again computed by using conventional distance after KSC found all centers. The potential of the cluster centers found by WMSC, KSC, SC and MC was plotted in Fig. 2-4 for ArtData, Iris and Pima respectively.

From Fig. 2, we could see the centers found by WMSC have higher potential than the ones found by the other three algorithms for ArtData. In Fig. 3, the first two centers with the highest potential were found by WMSC and the third one with the highest potential was found by KSC. In Fig. 4, the results of MC were not given due to memory and speed limitations. When r_a was set the same value for WMSC and SC for Pima dataset, WMSC found three centers and SC found six. The same number of cluster centers found by WMSC was adopted in KSC. The first and second centers with the highest potential were both found by WMSC and the third by KSC. In general, the cluster centers found by WMSC had higher potential than the ones found by SC for all three datasets.

Three validity indexes, including Davies-Bouldin (DB) index (Davies and Bouldin, 1979), Dunn's index (Bezdek and Pal, 1998) and Dehuri's index (Dehuri *et al.*, 2006) were used to evaluate the clustering algorithms. DB index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. A smaller value of DB index indicates a better clustering result. Dunn's index

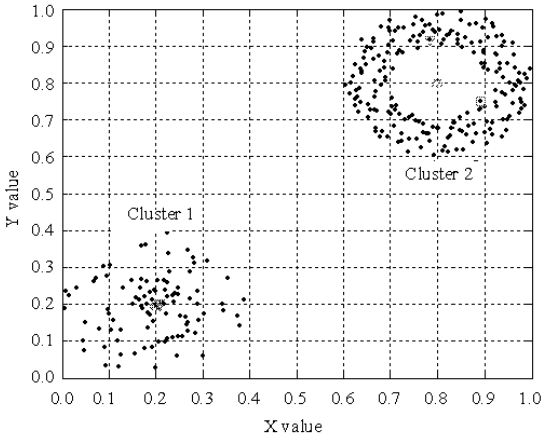

Fig. 1: The distribution of cluster 1 and cluster 2 and their cluster centers found by WMSC, KSC, SC and MC for ArtData. (○) the cluster center found by WMSC, (△) the cluster center found by KSC, (□) the cluster center found by SC, (▽) the cluster center found by MC

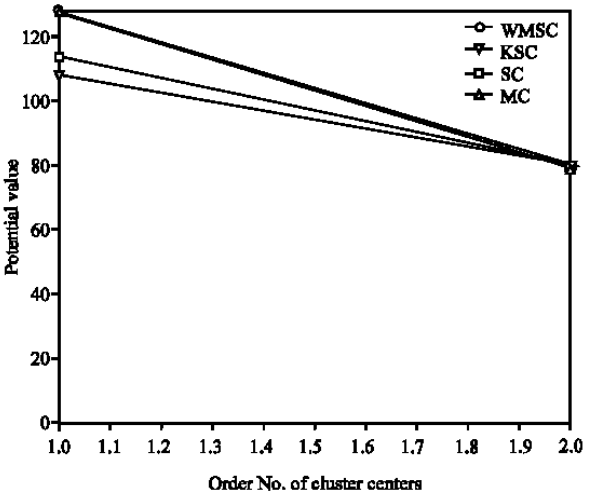

Fig. 2: The potential of cluster centers found by WMSC, KSC, SC and MC against the order number of cluster centers for ArtData

is designed to identify sets of clusters that are compact and well-separated. The greater value of Dunn' index, the better clustering result. Dehuri's index is the percentage of samples correctly classified. A greater value of Dehuri's index indicates a higher clustering accuracy. Five algorithms, fuzzy c-means and the four methods referred above, were used to demonstrate the effectiveness of our proposed algorithm. In the experiments, fuzzy c-means was run 30 times because

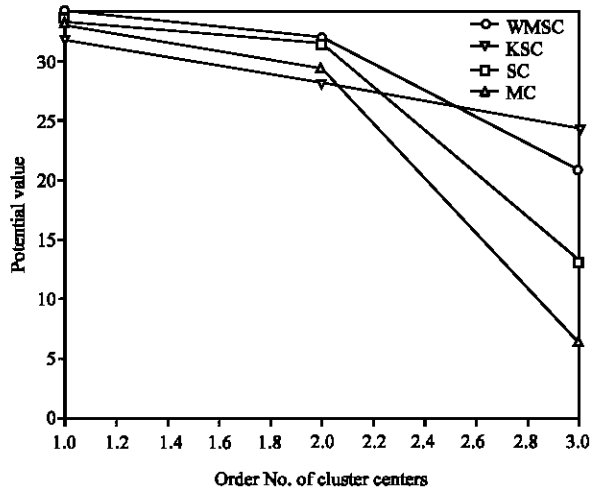

Fig. 3: The potential of cluster centers found by WMSC, KSC, SC and MC against the order number of cluster centers for Iris

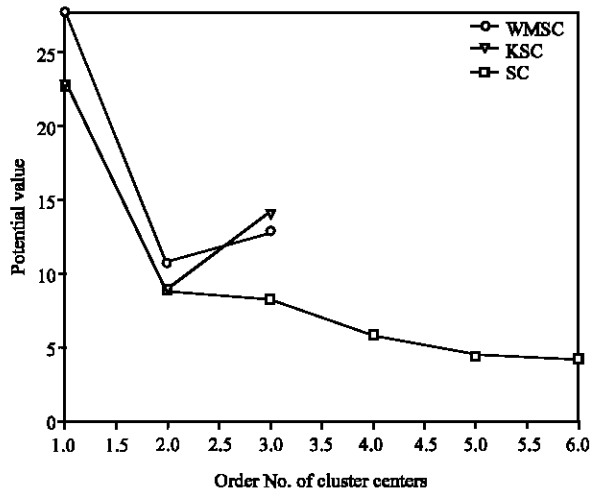

Fig. 4: The potential of cluster centers found by WMSC, KSC and SC against the order number of cluster centers for Pima

different results were obtained in different run and the other four algorithms were run 1 time because the determinable results were obtained for one dataset by one method. The parameters of fuzzy c-means and KSC were set the same as in (Kim *et al*., 2005). Table 1-3 list the values of three indexes of five algorithms for ArtData, Iris and Pima accordingly.

In Table 1, all algorithms obtained the same value of each of three indexes for ArtData. It is because the data has two compact and well-separated clusters. Even different cluster centers were found by different algorithms, the data was grouped into the same two

Table 1: The values of three indexes of each algorithm for ArtData dataset

Method	DB	Dunn's	Dehuri's
Fuzzy c-means	0.4370	1.2496	1
SC	0.4370	1.2496	1
MC	0.4370	1.2496	1
KSC	0.4370	1.2496	1
WMSC	0.4370	1.2496	1

Table 2: The values of three indexes of each algorithm for Iris dataset

Method	DB	Dunn's	Dehuri's
Fuzzy c-means	1.5915	0.0346	0.8027
SC	1.5905	0.0899	0.9730
MC	1.5645	0.1041	1
KSC	1.6191	0.0347	0.8696
WMSC	1.5222	0.1041	1

Table 3: The values of three indexes of each algorithm for Pima dataset

Method	DB	Dunn's	Dehuri's
Fuzzy c-means	8.1278	0.0461	0.5239
SC	49.6610	0.0528	0.6935
MC	-	-	-
KSC	8.0532	0.0446	0.5668
WMSC	7.4078	0.0593	0.5238

clusters by five algorithms. In Table 2, the best values of DB index, Dunn's index and Dehuri's index were all achieved by WMSC. The same values of Dehuri's index and Dunn's index were also achieved by MC. In Table 3, WMSC achieved the best values of DB index and Dunn's index but the worst value of Dehuri's index. A slightly higher value of Dehuri's index was obtained by Fuzzy c-means and the highest value by SC. In general, WMSC performed better than other algorithms on three validity indexes.

CONCLUSIONS AND FUTURE WORK

This study presents a weighted mean subtractive clustering algorithm in which cluster centers are derived by weighted mean method. Comparative experiments were executed among weighted mean subtractive clustering, fuzzy c-means, kernel-based subtractive clustering, conventional subtractive clustering and mountain clustering on three datasets. The experimental results indicate weighted mean subtractive clustering can create new centers with higher potential. Weighted mean subtractive clustering is superior to other algorithms in some aspects.

The kernel-based subtractive clustering applies kernel technique (Kim *et al*., 2005) to cluster data that is linearly non-separable in the original space into homogeneous groups in the transformed high dimensional space. The modified mountain clustering algorithm (Yang and Wu, 2005) uses new mountain function, which is convenient for automatically estimating the parameters in accordance with the structure of the data and also the number of clusters. Our method

specializes in creating new centers which are not included in original dataset. Each method improves subtractive clustering by a special way. It may be a better choice to introduce one's advantage to another method. The following ideas are involved in future work: (1) introducing kernel-induced distance instead of the conventional distance when calculating the potential value of data point in weighted mean subtractive clustering; (2) trying modified mountain function to compute potential and revised mountain function to update potential (Yang and Wu, 2005) in weighted mean subtractive clustering and (3) comparing weighted mean subtractive clustering in which different weighted mean methods are used to derive new cluster centers.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions to improve the paper. This work was supported by the National Basic Research Program of China (973 Program) under Grant No.2004CB719401.

REFERENCES

Bezdek, J.C. and N.R. Pal, 1998. Some new indexes of cluster validity. IEEE. Trans. Syst. Man Cyber. Part B, 28 (3): 301-315.

Bezdek, J.C. *et al*., 1999. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academy Publishers, Boston.

Blake, C.L. and C.J. Merz, 1998. UCI Repository of Machine Learning Databases. (http://www.ics.uci.edu/~mlearn/MLRepository.html).

Chiu, S.L., 1994. Fuzzy model identification based on cluster estimation. J. Intel. Fuzzy Syst., 2 (3): 267-278.

Davies, D.L. and D.W. Bouldin, 1979. A cluster separation measure. IEEE. Trans. Pattern Anal. Mach. Intel., 1 (4): 224-227.

Dehuri, S., C. Mohapatra, A. Ghosh and R. Mall, 2006. A comparative study of clustering algorithms. Inform. Technol. J., 5 (3): 547-553.

Kim, D.W., K.Y. Lee, D. Lee and K.H. Lee, 2005. A Kernel-based subtractive clustering method. Pattern Recog. Lett., 26 (7): 879-891.

Li, H., L.Y. Shen and P.E.D. Love, 1999. ANN-based mark-up estimation system with self-explanatory capacities. J. Construct. Eng. Manage., 125 (3): 185-189.

Nikhil, R.P. and D. Chakraborty, 2000. Mountain and subtractive clustering method: Improvements and generalizations. Int. J. Intel. Syst., 15 (4): 329-341.

Yager, R.R. and D.P. Filev, 1994. Approximate clustering via the mountain method. IEEE. Trans. Syst. Man Cyber., 24 (8): 1279-1284.

Yang, M.S. and K.L. Wu, 2005. A modified mountain clustering Algorithm. Pattern Anal. Applied, 8 (1): 125-138.