An Improved CURE Algorithm

Mingjuan Cai, Yongquan Liang

An improved CURE algorithm

Mingjuan Cai [1], Yongquan Liang [2]

[1]College of Computer Science and Engineering and Shandong Province Key Laboratory of Wisdom Mine Information Technology,
Shandong University of Science and Technology, Qingdao 266590, China
[2]College of Computer Science and Engineering,
Shandong University of Science and Technology, Qingdao 266590, China
lyq@sdust.edu.cn

ABSTRACT. CURE algorithm is an efficient hierarchical clustering algorithm for large data sets. This paper presents an improved CURE algorithm, named ISE-RS-CURE. The algorithm adopts a sample extraction algorithm combined with statistical ideas, which can reasonably select sample points according to different data densities and can improve the representation of sample sets. When the sample set is extracted, the data set is divided at the same time, which can help to reduce the time consumption in the non-sample set allocation process. A selection strategy based on partition influence factor is proposed for the selection of representative points, which comprehensively considers the overall correlation between the data in the region where a representative point is located, so as to improve the rationality of the representative points. Experiments show that the improved CURE algorithm proposed in this paper can ensure the accuracy of the clustering results and can also improve the operating efficiency.

Keywords: clustering algorithm, CURE algorithm, sampling, representative point

1. Introduction

Clustering is a basic task of data mining and is also an unsupervised learning process. The goal of clustering is to gather the *n* objects in a given d-dimensional space into k clusters according to a specific data metrics, and to make the objects in a cluster have a maximum similarity degree, while the objects between the clusters have a minimum similarity degree[1]. The existing clustering methods are mainly divided into four categories (Berkhin,2006): partitioning method, hierarchical method, density method and model-based method[2].

Each algorithms has its own advantages and disadvantages when solving problems.The CURE (Clustering Using REpresentatives) algorithm is an agglomerative hierarchical clustering method[3-4]. In recent years, researchers have proposed many novel algorithms from different angles to improve the CURE algorithm. For example: Kang Weixian et al. introduced the CURE algorithm in detail in the literature [5] and proposed an improved algorithm. Shen Jie proposed an

improved algorithm K-CURE[6]. Wu Heng et al. proposed an improved algorithm RTCURE[7] based on information entropy for measuring inter-class distances. Inspired by these, a new improved algorithm ISE-RS-CURE (CURE algorithm based on improved sample extraction and representative point selection) is proposed to optimize the sample extraction and representative point selection process of CURE algorithm in this paper. Experiments show the algorithm has better performances as well as low time complexity.

2. Related work

The CURE algorithm can find arbitrarily shaped clusters and is insensitive to noise points. It is different from the traditional clustering algorithm using a single particle or object to represent a class, while choosing a fixed number of representative points to represent a class[8]. This intermediate strategy[9] based on particle and representative object method can detect clusters of different shapes and sizes, and is more robust to noise, overcoming the problem that most clustering algorithms are either only good at dealing with clusters of spherical and similar sizes, or they are relatively fragile when dealing with isolated points.

The CURE algorithm includes both the hierarchical part and the divided part, which overcomes the disadvantage of using a single clustering center tend to discover spherical clusters. A large number of experiments and experiments have proved that the CURE algorithm is effective. Under normal circumstances, the value of the contraction factor is between 0.2 and 0.7, and the number of points larger than 10 can get the correct clustering result[10].

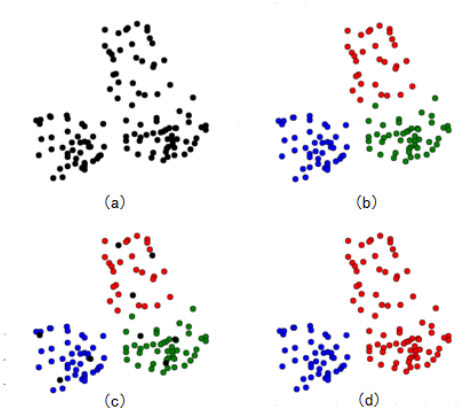

Fig. 1. The basic steps of CURE algorithm

Figure1 shows the basic steps of the CURE algorithm. The data set is listed in Figure1(a). For large-scale data, it is usually obtain a sample set by random sampling. Figure1(b) shows three clusters of current clustering, expressed in three colors respectively. Figure1(c) shows the representative points selected from each cluster are shown by the black point. Figure1(d) shows after "shrinkage" of the representative

points, the two clusters with highest similarities are merged. Then, reselect the representative points of the new cluster and repeat the two processes of "shrinking" the representative points and merging the clusters until the preset number of clusters is reached. If a class grows slowly during clustering or contains very little data at the end of the cluster, it is treated as noise.

3. Improved CURE algorithm

3.1. Sample extraction

The traditional CURE algorithm uses random sampling to obtain sample sets, which is ideal for uniformly distributed data sets; however it may result in incorrect clustering results for non-uniformly distributed data[11]. This paper uses a sample extraction algorithm, which combines statistical ideas to ensure that the data set can be processed more efficiently and the sample set can describe the dataset more accurately. Before introducing the algorithm, we need to establish the mathematical statistics model of the data set first, and then use the merge decision criterion based on independent finite difference inequalities to complete the data extraction.

Definition 1 $A = \{A_1, A_2, \dots, A_n\}$ is the feature set of data set. Any feature A_i can be represented by m independent random variables, where $1 \le i \le n$, $m > 0$. Suppose the range of any feature is $[L_i, R_i]$, and after normalization, the range of the feature is changed to $[0, h_i]$, where $h_i \le 1$. Therefore, the range of values for m independent random variables should also be $[0, h_i]$, and the range of each independent random variable is $[0, h_i/m]$.

When all data are represented by multiple sets of random variables, the statistical model of the data set is established. In the statistical model, sampling of data points is independent of each other, and no distribution of m random variables is required. The criterion for clustering merged data points is mainly derived from the independent finite difference inequality.

Theorem 1 Let $X = (X_1, X_2, \dots, X_m)$ be a set of independent random variables. The range of X_i is $l_i (1 \le i \le m)$. Suppose there is a function $\boldsymbol{F}$ defined in $\prod_i X_i$. When the variables X and X' only differ in the k-th condition, $|F(X) - F(X')| \le r_k$ is satisfied. There has

$$P\big(F(X) - E\big(F(X)\big) \ge \tau\big) \le exp\left(-2\tau^2 / \sum_k (r_k)^2\right) \tag{1}$$

Definition 2 Suppose q is the number of features of data items in the data set, and any feature can be represented by m independent random variables. If there exist a class C, the number of data points in class C is represented by $|C|$, the expectation of expected sum of m random variables of related data points in class C is expressed by $E(C)$, (C_1, C_2) is a combination of classes, C_{1c} and C_{2c} represent the core points of classes C_1 and C_2. The merging criterion is formula[2], where $0 < \delta \le 1$.

$$|(C_{1c} - C_{2c}) - E(C_{1c} - C_{2c})| < \sqrt{\frac{1}{2mq}\left(\frac{|C_1| + |C_2|}{|C_1||C_2|}\right) ln\, \delta^{-1}} \tag{2}$$

Proof: Firstly, prove

$$|(C_{1c} - C_{2c}) - E(C_{1c} - C_{2c})| \geq \sqrt{\frac{1}{2mq}\left(\frac{|C_1| + |C_2|}{|C_1||C_2|}\right) ln\, \delta^{-1}} \tag{3}$$

conforms to theorem 1.

Let $\tau = \sqrt{\frac{1}{2mq}\left(\frac{|C_1|+|C_2|}{|C_1||C_2|}\right) ln\, \delta^{-1}}$, and there have

$$max\left(\sum\nolimits_k (r_k)^2\right) = mq|C_1| \cdot max\left(r_{C_1}\right)^2 + mq|C_2| \cdot max\left(r_{C_2}\right)^2$$

$$= \frac{1}{mq}\left(\frac{|C_1| + |C_2|}{|C_1||C_2|}\right) \tag{4}$$

then

$$P(|(C_{1c} - C_{2c}) - E(C_{1c} - C_{2c})| \geq \tau) \leq exp\left(-2\tau^2 / \sum_k (r_k)^2\right)$$

$$\leq exp\left(-2\tau^2 / max(\sum_k (r_k)^2)\right) = \delta \tag{5}$$

From formula[4] we can see that the formula[3] conforms to theorem 1. So when δ approaches 0 (δ=0.0001 in this paper), the probability of $|(C_{1c} - C_{2c}) - E(C_{1c} - C_{2c})| < \tau$ is close to 1. □

Let $b(C_1, C_2) = \sqrt{\frac{1}{2mq}\left(\frac{|C_1|+|C_2|}{|C_1||C_2|}\right) ln\, \delta^{-1}}$, and if C_1 and C_2 belong to the same class, then $E(C_{1c} - C_{2c}) = 0$. It can be concluded that if the class combination (C_1, C_2) satisfies inequality$|(C_{1c} - C_{2c})| < b(C_1, C_2)$, it can be merged. Therefore, the merging criterion is defined as: For the class combination (C_1, C_2), when each feature satisfies$|C_{1c} - C_{2c}| < b(C_1, C_2)$, the C_1 and C_2 are merged; otherwise the merging is not performed.

3.2. Representative point selection

The reasonable selection of representative points can directly affect the quality of clustering result[12]. A representative point selection method based on the distance influence factor is proposed in the literature[13], the appropriate representative point can be selected in different regions of the cluster according to the density distribution of the cluster, but the distance influence factor ignoring the influence of the entire district where the representative point is located. In this paper, a representative point selection method based on Partition Influence Weight (PIW) is proposed, which takes into account the influence of the partition where the representative point is located in the cluster, can effectively eliminates noise points and the appropriate representative points can be selected according to the density distribution of the cluster.

Definition 2 $C = \{d_1, d_2, \cdots, d_n\}$ is a cluster in data set, the representative point in the cluster is d_{ri} ($0 < i \leq \sqrt{|C|}$), and C_c is the core point. Under the minimum distance between the data point and the representative point, a cluster partition

$\{C_1, C_2, \cdots\}$ is obtained. Each of the segments C_i has a one-to-one correspondence with the d_{ri}. The representative point d_{ri} has a weight of

$$PIW(d_{ri}) = \frac{n}{m} \cdot \frac{\sum_{j=1}^{m} d(d_j, C_c)}{\sum_{i=1}^{n} d(d_i, C_c)} \cdot dist(d_{ri}, C_c) \quad (6)$$

Among them, n is the number of data items in the cluster C, m is the number of data items in the block C_i, and $d(d_i, C_c)$ is the distance from the core point C_c to each data item in the cluster C, and $d(d_j, C_c)$ is the distance from the core points C_c to the data items in the block C_i.

In the process of selecting the representative point, the $PIW(d_{ri})$ value is compared with the threshold value η, and the adjustment of the representative point selection is performed until the proper representative point is selected. The initial threshold is set as $\eta = \frac{n}{5m} \cdot \frac{\sum_{j=1}^{m} d(d_j, C_c)}{\sum_{i=1}^{n} d(d_i, C_c)} \cdot dist(d_{ri}, C_c)$.

3.3. The ISE-RS-CURE algorithm

The ISE-RS-CURE algorithm firstly uses the sample extraction algorithm based on statistical ideas is used to extract samples, in which the merge criterion is used instead of the distance threshold. Removing the representative points with count less than 2 can effectively eliminate the noise points and reduce the interference of noise points and abnormal points. This sample extraction algorithm carries out a rough partition of the sample set while obtaining the sample set. This information can be used to complete the final clustering result during the stage of labeling the non-sample data, which can effectively improve the efficiency of the algorithm. The partition strategy in the representative point selection algorithm guarantees the distribution of representative points. The partition influence factor can be used to continuously adjust the selection of representative points. The noise points can be eliminated so that the representative point set can better reflect the shape size and density information of the data set, and avoid the phenomenon that the representative point gathers and the noise point becomes the representative point. The pseudo code for the ISE-RS-CURE algorithm is:

Algorithm 1. ISE-RS-CURE algorithm

Input: data set *D*; the number of independent random variables *m*
Output: Clustering results
//Phase 1: Get a sample set
1.Initialize the sample set *S*
2.for each data item *x* in data set *D* do
3. for each sample point *s* in sample set *S* do
4. if *x* satisfy: $s_i \in S$, $s_i$located C_i, for all features there is $|c_{ik} - x_k| < b(C_i, x)$, then $C_i = C_i \cup \{x\}, D = D - \{x\}$
5. else $S = S \cup \{x\}, D = D - \{x\}$
6. end for
7.end for

8. Delete the s_i from S, where $|C_i| < 2$, and mark s_i as noise point
//Phase 2: Clustering stage
9. Initialize the representative set R
10.Each item in sample set S is treated as a class and aggregated hierarchical clustering
11. for each cluster C_i do
12. random select d_i, where $d_i \in C_i$, if $d(d_i, C_c) = min\left(max\left(d(d_i, C_c)\right)\right)$,let $C_c \leftarrow d_i, R_i = R_i \cup \{d_i\}$. Cycle $\sqrt{|C_i|}$ times.
13. divide the cluster C_i according to R_i, and get $\left\{C_{i1}, C_{i2}, \cdots, C_{i\sqrt{|C_i|}}\right\}$
14. calculate the $PIW\left(d_{rij}\right)$ for each partition C_{ij}
15. if $PIW\left(d_{rij}\right) \leq \eta$ then
16. Delete d_{rij}, select a new representative point and update R_i
17.end for
18."shrink" the representative point to the cluster center point
19.Merge the two clusters with the highest similarity and return to **step 11** until the number of target clusters is reached
20.Assigning non-sample data items to different clusters in data sets

In the Step (16) the selection strategy of the new representative point is: in the cluster where the representative point d_{rij} is located, a data item that has not been selected as a representative point is randomly selected as a new representative point, and the value of $PIW\left(d_{rij}\right)$ is recalculated until all the weights of the representative points are all greater than η. The sample extraction algorithm requires only one scan of the data set. Therefore, for a given data set of size *n*, the time complexity is only *O(n)*, and the advantage is obvious when dealing with large-scale data sets. The representative point selection algorithm needs to calculate the distance metric matrix of the sample set. The time complexity is $O(m^2)$, and *m* is the number of sample set data points. Thus, the time complexity of the ISE-RS-CURE algorithm is $O(n + m^2)$ and the space complexity is *O(n)*.

4. Experiment analysis

In order to test the actual clustering performance of the improved CURE algorithm presented in this paper, the experimental part compares the traditional CURE algorithm and the ISE-RS-CURE algorithm by using the synthetic data set and the real data set respectively.

4.1. Experiments on Synthetic data sets

Experiment 1: In order to compare the ISE-RS-CURE algorithm and the traditional CURE algorithm in the accuracy of the clustering results, the experiment uses an synthetic data set D1 (as shown in Figure 2). The data distribution of data set D1 is complex, which contains 6 clusters with different shape size and large location

difference, and dense noise points with cosine distribution and a large number of scattered noise points. The total number of data points is 8000 and the number of attributes is 2. For ease of discussion, the six major clusters clustered are numbered as ① to ⑥ respectively.

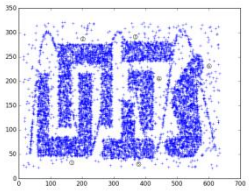

Fig.2. Data set D1

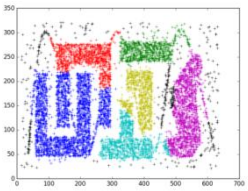

Fig.3.CURE clustering results

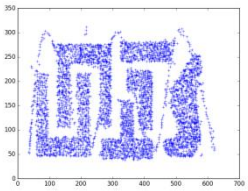

Fig.4.ISE-RS-CURE sample set

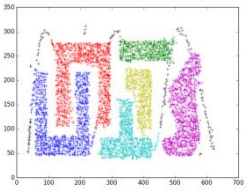

Fig.5. Sample set clustering results

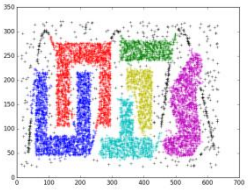

Fig.6.ISE-RS-CURE clustering results

Figure 3 shows the clustering results of the traditional CURE algorithm, it can be seen that due to the influence of cosine noise points, cluster ① and cluster ② are different from ideal clusters, and ④⑤⑥ is also disturbed to varying degrees. Figure 4 is a sample set obtained by the sampling algorithm in ISE-RS-CURE algorithm. It can be seen that the sample set conforms to the distribution of the original data set, and the noise data with lower density is filtered out. The sample set has a higher representative. Figure 5 shows the clustering result of the sample set. Since the noise points have less interference, the correct clustering results can be displayed for the six major clusters. Figure 6 shows the final clustering result of the ISE-RS-CURE algorithm. Compared with the traditional CURE algorithm, the ISE-RS-CURE can get clustering results that are closer to the real cluster, so it has significantly improved the clustering accuracy.

Experiment 2: In order to compare the run-time performance of the traditional CURE algorithm and the ISE-RS-CURE algorithm, this experiment uses an synthetic data set D2 (as shown in Figure 7). Data set D2 is a data set of 3 round clusters composed of random numbers. Because the data set is distributed evenly and there is no noise point, the traditional CURE algorithm and the improved CURE algorithm can both get the correct clustering results. In the experiment, the traditional CURE algorithm uses random sampling to obtain the same size sample set. The number of data points is increased from 5000 to 25000, and comparing the running time of the two algorithms.

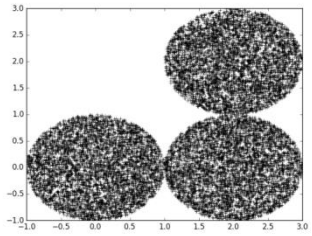

Fig. 7. Data set D2

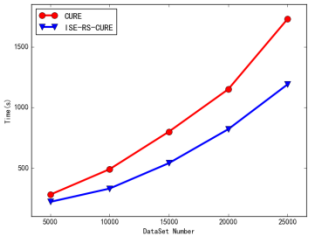

Fig. 8. Time contrast of two algorithms

It can be seen from Figure8 that with the increase of data number, the execution time of the traditional CURE algorithm increases exponentially. Although the execution time of the improved algorithm also increases exponentially, it obviously slows down; the execution time of the ISE-RS-CURE algorithm is less than the execution time of the traditional algorithm, and as the data volume increases, the gap gradually increases. Therefore, the ISE-RS-CURE algorithm has an improvement in execution time compared with the traditional CURE algorithm.

4.2. Experiments on the real data sets

This experiment selects 5 data sets in the UCI machine learning database[14], tests the traditional CURE algorithm and the ISE-RS-CURE algorithm, and evaluates it with 3 evaluation criteria. The basic attributes of the dataset are display in Table1.

Table 1. Data set table

Data set	Size	Attribute	Cluster
Wine	178	13	3
Seeds	210	7	3
Ionosphere	351	34	2
Waveform	5000	31	3
Segmentation	2310	19	7

The clustering results are compared and analyzed with 3 evaluation indexes: CA, ARI and NMI[15]. The results of each data set and corresponding evaluation indicators are recorded in Table2.

Table 2. Comparison of results of evaluation indicators on UCI datasets

Data set	Algorithm	CA	ARI	NMI
Wine	CURE	0.7394	0.5271	0.7058
	ISE-RS-CURE	**0.7401**	**0.5389**	**0.7125**
Seeds	CURE	0.6910	0.5936	0.6773
	ISE-RS-CURE	**0.7028**	**0.6192**	**0.6854**
Ionosphere	CURE	0.6283	0.7492	0.6787
	ISE-RS-CURE	**0.6398**	**0.7635**	**0.7021**
Waveform	CURE	0.5862	0.3368	0.4920
	ISE-RS-CURE	**0.6750**	**0.4068**	**0.5473**

Segmentation	CURE	0.6449	0.4839	0.6281
	ISE-RS-CURE	**0.7061**	**0.5325**	**0.6657**

From Table2, we can see that for the five different data sets, the ISE-RS-CURE algorithm performs better than the traditional CURE algorithm on the 3 clustering evaluation indexes of CA, ARI and NMI. Especially in the large amount of data and attribute multiple data sets, experiments show that the ISE-RS-CURE algorithm has more obvious advantages, which further proves the effectiveness of the ISE-RS-CURE algorithm in clustering.

5. Conclusion

This paper proposes an ISE-RS-CURE algorithm for the disadvantages of the traditional CURE algorithm, which sample set can't reflect the data distribution and Representative points are less representativeness. The ISE-RS-CURE algorithm uses a combined decision criterion based on statistical ideas for sample extraction and the representation point selection strategy based on the partition influence weight. It can not only obtain the set of samples which conform to the data set distribution, but also can gather arbitrary shape clusters and have better robustness. Experimental results show that the ISE-RS-CURE algorithm can effectively improve the accuracy and efficiency of the algorithm.

References

1. Han Jiawei, Kamber M. Data mining: concepts and techniques. 3nd ed. Beijing: China Machine Press(2012).
2. Niu, Z-H., Fan, J-C., Liu, W-H., Tang, L. and Tang, S. CDNASA: clustering data with noise and arbitrary shape, Int. J. Wireless and Mobile Computing, pp.100–111, Vol. 11, No. 2 (2016).
3. Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. In: ACM SIGMOD International Conference on Management of Data. ACM,73-84(1998).
4. Guha S, Rastogi R, Shim K, et al. CURE : An Efficient Clustering Algorithm for Large Databases. Information Systems, 26(1):35-58(2001).
5. Kang Wei-xian, Ye De-qian. Study of CURE Based Clustering Algorithm. In: 18th China Conference on Computer Technology and Applications (CACIS) (Volume One). Computer Technology and Application Progress, Hefei: China University of Science and Technology Press, 132-135(2007).
6. SHEN Jie, ZHAO Lei, YANG Ji-wen, et al. Hierarchical clustering algorithm based on partition. Computer Engineering and Applications, 43(31):175-177(2007).
7. Wu Heng, Li Wenjie, Jiang Min. Modified CURE clustering algorithm based on entropy. Computer application research, 34(08):2303-2305(2017).
8. WANG Yin-tong, WANG Jian-dong, CHEN Hai-yan, XU Tao, Sun Bo. An Algorithm for Approximate Binary Hierarchical Clustering Using Representatives. Mini-Micro Computer Systems, 36(02):215-219(2015).
9. J Fray B J, Dueck D. Clustering by Passing Messages Between Data Points. Science, 315(5814): 972-976(2007).

10. JIA Rui-yu, GENG Jin-wei, NING Zai-zao, et al. Fast clustering algorithm based on representative points. Computer Engineering and Applications, 46(33):121-123+126(2010).
11. Zhao Y. Research on user clustering algorithm based on CURE. Computer Engineering & Applications, 11(1):457-465(2012).
12. Shao X, Wei C. Improved CURE algorithm and application of clustering for large-scale data. In: International Symposium on It in Medicine and Education. IEEE, 305-308(2012).
13. SHI Nian-yun, ZHANG Jin-ming, CHU Xi. CURE Algorithm-based Inspection of Duplicated Records. Computer Engineering, 35(05):56-58(2009).
14. Lichman M. UCI machine learning repository [EB/OL](2013), http:/ /archive.ics.uci.edu /ml. 2018/02/24.
15. LU Pengli, WANG Zudong. Density-sensitive hierarchical clustering algorithm. Computer Engineering and Applications, 50(04):190-195(2014).