# A Dynamic Auto-Stopped Clustering Algorithm Based on Outlier Information

**Tian-yang Lv[1, 2], Shao-bin Huang[1], Wan-li Zuo[2], Zheng-xuan Wang[2]**

1: College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
2: College of Computer Science and Technology, Jilin University, Changchun 130012, China

## Abstract

Traditional hierarchical and partitioning clustering algorithms usually require the user-specified number of final clusters *k*. Meanwhile, they lack the mechanism to handle outliers or regard outlier-detection as the byproduct of the clustering process by treating outliers as "noise". This paper takes a new perspective and complements classical approaches by regarding outliers as valuable information. The paper proposes an auto-stopped clustering algorithm CURED that is constructed on a new outlier-mining algorithm and the improved CURE algorithm. 3 datasets are adopted to evaluate CURED and the experimental results show its good performance in both outlier detection and clustering.

**Keywords**: Clustering, outlier detection, auto stop

## 1. Introduction

The hierarchical clustering algorithms, such as CURE [2] and the partitioning algorithm K-means usually require the user-specified number of final clusters *k*. However, it is very difficult to choose an appropriate *k*, because of lacking the valuable prior knowledge or the inconsistency between the prior knowledge and data's realistic distribution situation.

The clustering algorithms are also short at outlier detection. Some algorithms prune the small clusters or the clusters growing very slowly as outliers. However, this is treating outlier-detection as the byproduct of clustering and cannot perform the outlier detection task well. It is also a waste of valuable information in deleting outliers as "noise".

To make up these drawbacks, the paper proposes a new strategy that combines outlier detection with clustering. It is based on the following observation: the distances among data or clusters not only show their similarity degree, but also demonstrate the dissimilarity. And with the progressing of clustering, the dissimilarity $D(C_{NN\text{-}A}, C_{NN\text{-}B})$ between the two most similar clusters $C_{NN\text{-}A}$ and $C_{NN\text{-}B}$ at present is increasing. Therefore, the clustering should stop at the moment that $C_{NN\text{-}A}$ and $C_{NN\text{-}B}$ are so diverse from each other.

To realize this strategy, after introducing related works in section 1, a new outlier detection method is proposed in section 2; section 3 introduces the new algorithm CURED, which is constructed on the outlier detection method and the improved CURE; section 4 gives the experimental result; section 5 summarizes the paper.

Table 1 Important Notation

| | |
|---|---|
| $N$ | Total number of Data |
| $M$ | Dimensionality of Data |
| $k$ | Number of Final Clusters |
| $C_i$ | The $i$ th Cluster |
| $n_i$ | Size of $C_i$ |
| $D(C_i, C_j)$ | The distance between $C_i$ and $C_j$ |

### 1.1. Related works

Comparing with other algorithms, CURE employs several novel approaches: using *r* data (*representative*) to represent a cluster; shrinking the *representative*s towards the cluster's centroid by a fraction $\alpha$; deleting the small clusters or slowly growing clusters as outliers. Its overview is:

(1) Pre-decide parameters $k, \alpha, r$, each input data is treated as a cluster; (2) Find the nearest neighbor of each cluster and the distance *D* between clusters equals the minimum distance among all their shrunk *representative*s; (3) Merge the closest pair of clusters and determine the *representative*s of the new-born cluster; if more than *k* clusters remain, go to (2), else stop.

The way to choose the *representative*s is: if $r \geq n_i$, all data of $C_i$ are *representative*s; otherwise, the first *representative* is the farthest one from $C_i$'s centroid and data farthest from the previous chosen *representative* is selected as the next *representative*, do this iteratively till *r representative*s are decided.

However, there are several shortcomings of CURE: (1) needing the user-specified number *k* of the result clusters; (2) interfered by the outliers before they are pruned and the information implied

by outliers is wasted; (3) a cluster's density is not taken into consideration in the merging decision.

Some researches attend to make clustering algorithm optimally estimate $k$. [1] adopts the global criterion function $f$: stop clustering once $f$ is optimized. Its shortcomings are: prone to fall into local optimization; needing new parameter; difficult in choosing the appropriate criterion. [3] proposes a new hierarchical clustering method based on dissimilarity. But it is also short at outlier detection.

## 2. A New Outlier Detection Algorithm

The basic idea of distance-based outlier detection method is: if the distances of data $a$ and most other data are larger than the threshold $D_{\text{out}}$, $a$ is an outlier. This section proposes to decide the appropriate threshold $D_{\text{out}}$ according to the even distribution pattern of data and consider data's local distribution feature during outlier detection.

The even distribution pattern is a very useful reference, since clusters and outliers exist only if the real-life data distribute unevenly. And when data distribute evenly in the vector space $S$, the distances $\overline{D}_{NN}$ of each data and its nearest neighbor are the same. To compute $\overline{D}_{NN}$, $S$ is equally divided into $N$ grids with only one data in grid's center. And the length of $i$ th-dimensional edge of a grid is $(a_{\max}^{(i)} - a_{\min}^{(i)}) / \sqrt[M]{N}$, where $a_{\max}^{(i)}$ and $a_{\min}^{(i)}$ is the maximum and the minimum of all data's $i$ th-dimension. Thus:

$$\overline{D}_{NN} = \sqrt{\sum_{i=1}^{M} ((a_{\max}^{(i)} - a_{\min}^{(i)}) / \sqrt[M]{N})^2} \qquad (1)$$

Parameter $\beta$ is adopted to describe the diversity degree of the realistic distribution situation from the even pattern and $D_{\text{out}} = \overline{D}_{NN} / \beta$.

Factor $\xi$ is adopted to evaluate the local distribution feature of a data,. For data $a$, $\xi(a) = D_{\text{NN}}(a) / D_{\text{NN}}(b)$, where $D_{\text{NN}}(a)$ is the distance between $a$ and its nearest-neighbor and so is $D_{\text{NN}}(b)$. $\xi(a)$ shows the isolation degree of $a$ from its neighbors. The special method should be used for the very similar or duplicate data. For instance, $b$ is $a$'s nearest neighbor and $b$ and $c$ are very similar, which means $D_{\text{NN}}(b) \to 0$, thus $a$ would be regarded as an outlier no matter what value $D_{\text{NN}}(a)$ is. To avoid this, adjust the equation for $\xi(a)$ as follows:

$$\xi(a) = \begin{cases} D_{NN}(a) / D_{NN}(b) & if\ (D_{NN}(b) > 10^{-4}) \\ 1 & else \end{cases} \qquad (2)$$

And the outlier evaluation criterion is:

Data $a$ is an outlier, if

$$D_{NN}(a) * \xi(a) > \left(\frac{\sqrt{\sum_{i=1}^{M} ((a_{\max}^{(i)} - a_{\min}^{(i)}) / \sqrt[M]{N})^2}}{\beta}\right) \qquad (3)$$

An interesting phenomenon can be observed in the outlier detection process: $n_{\text{out}}$ increases much faster with further decreasing of $D_{\text{out}}$ after all outliers are detected. It is because outliers are extremely far away from the others. Therefore, we propose a method to decide the suitable value of $\beta$:

Let $\beta_{\text{Step}}$ equals $\beta$ when the first outlier is detected. Then, $\beta_{Step} = D_{NN}(a_{far}) \times \xi(a_{far}) / \overline{D}_{NN}$, where $D_{NN}(a_{far}) \times \xi(a_{far}) \geq D_{NN}(b) \times \xi(b)$ for any $b$. Observe the increasing speed $V$ of $n_{out}$ with the increase of $\beta$, where $\beta = l \times \beta_{Step}$ and $l = \{1, 2, \ldots\}$. And $\beta = (l_i - 1) \times \beta_{Step}$, when $V$ reaches its first peak at $l_i$.

Since $a_{\max}$ and $a_{\min}$ can be decided in data reading and $\beta_{\text{Step}}$ is a byproduct in computing the nearest neighbor of all data, they are the input for the outlier detection process.

## 3. The Auto-stopped Clustering Algorithm

The dynamic auto-stopped clustering algorithm CURED is founded on the new outlier detection algorithm and the improved CURE algorithm. Fig. 1 is the overview of CURED. State the improvements made on CURE as follows.

### 3.1. Outlier detection

To compensate CURE in outlier mining, CURED performs outlier detection in two phases:

(1) Detecting outliers using the method of section 2 before clustering;

(2) Treating very small clusters in the clustering result as outliers.

The first phase reduces the interruptions coming from outliers for clustering. Therefore, it is sufficient to avoid the influence of the remaining outliers with $\alpha$'s value range [0.90, 1.00].

### 3.2. Distance between clusters

Clusters' density is taken into consideration in determining whether two clusters should be merged. Therefore, $D(C_i, C_j)$ is decided according to two factors: first, the distance $D_{\min}(C_i, C_j)$ of the nearest *representative*s coming from $C_i$ and $C_j$ respectively;

second, the factor $\delta$ measuring the change of cluster's density *Den* if $C_i$ and $C_j$ are merged.

The density of $C_i$ or $C_j$ approximately equals the average distances among its *representative*s. For the new-borne cluster $C_{new}$ created by merging $C_i$ and $C_j$, $Den(C_{new})=D_{min}(C_i,C_j)$. Then, $\delta(C_i)$ is defined as follows and so is $\delta(C_j)$:

$$\delta(C_i)=\begin{cases} Den(C_i)/Den(C_{New}) \\ \qquad if\,(Den(C_i)>Den(C_{New})) \\ Den(C_{New})/Den(C_i) \quad else \end{cases} \tag{4}$$

For the cluster with only one data, $D(C_i,C_j)=D_{min}(C_i,C_j)$. Therefore, the way to compute $D(C_i,C_j)$ is:

$$D(C_i,C_j)=\begin{cases} D_{Min}(C_i,C_j)\times(\delta(C_i)+\delta(C_j))/2 \\ \qquad if\,(n_i>1)\,\&\&\,(n_j>1) \\ D_{Min}(C_i,C_j) \quad else \end{cases} \tag{5}$$

Easy to prove that $(\delta(C_i)+\delta(C_j))/2\geq 1$, which means the bigger the difference between the density of $C_i$ or $C_j$ with that of $C_{new}$, the less possibility for $C_i$ and $C_j$ to be merged.

## 3.3. Automatic stop

Without user-specified *k*, it is necessary to extract the information from the processed data. As stated in former parts, it is a suitable opportunity to stop clustering if the clusters to be merged are too dissimilar.

We propose $D_{out}$ as the dissimilarity threshold to decide this opportunity for two reasons: (1) $D_{out}$ is used to detect outliers and the major characteristic of outliers is their dissimilarity from the others; (2) $D_{out}$ is decided according to the even distribution pattern and the existence of clusters shows the diversity from that pattern.

Therefore, the stop criterion for clustering is that: suppose $C_{NN\text{-}A}$ and $C_{NN\text{-}B}$ are the clusters to be merged, stop clustering if $D(C_{NN\text{-}A},C_{NN\text{-}B})>D_{out}$.

## 3.4. Complexity analysis

The complexity of CURE algorithm is $O(M*N^2)$. Since CURED is an algorithm based on CURE, we only analyze the complexity changes caused by each improvement. The complexity increases by $O(N)$ to scan outliers. It is $O(r^2)$ to compute a cluster's density. Since it is needed to compute the density of the new-born clusters ($N$-$k$) times, the complexity increases by $O(r^2*(N-k))$. And $k$ influences the complexity indirectly.

Take all these in total, the complexity increases by $O((N-k)*r^2+N)$. Since $r^2<N$ in most cases, the complexity of CURED equals that of CURE.

```
Algorithm CURED(α , r)
1. {   while (Not End)   //read all M-dimensional data
2.       {   read data a;
3.           decide vector a_max and a_min; }
4.       Treat each data as a cluster;
5.       Compute each cluster's nearest-neighbor;
6.       Determine the value of β_Step;
7.       Name the nearest pair at present as C_NN-A, C_NN-B;
8.       outlier (a_max, a_min, β_Step)   // detect outliers
9.       while ( D(C_NN-A, C_NN-B)≤D_out )
10.      {   Merge clusters C_NN-A and C_NN-B;
11.          Update C_NN-A and C_NN-B;   }
12.      Output the clustering result;
13.  }   //End of CURED
```

Figure 1 The Auto-stopped Clustering Algorithm CURED

# 4. Experiment and Analysis

Previous researches usually treat clustering and outlier-detection as separate topics and little effort has been taken to try to combine them. In contrast, the CURED algorithm makes clustering work together with outlier detection quite well: the outlier-detection algorithm excludes most disturbances of outliers for clustering and the outlier information is utilized in the determination of *k*.

## 4.1. Data sets and Criterion

The experiment adopts 3 data sets to compare CURED with CURE, ROCK and the algorithm of [3] named as Frozen algorithm.

Data set 1 is statistics of 330 NBA players [5] with 3 attributes. Data set 2 is the Zoo dataset with 101 records and 16 categorical attributes for each record [4]. Data set 3 consists of 193 sample images of 42 persons selected from the ORL Database and [7]. PCA method is applied to the images to extract 100-dimension feature data.

To measure the clustering results' quality, two criterions [6] *Entropy* and *Purity* are adopted. And the better the clustering result, the smaller is *Entropy* and the bigger is *Purity*.

$$Entropy=\sum_{i=1}^{k}\frac{n_i}{N}\left(-\frac{1}{\log q}\sum_{j=1}^{q}\frac{n_i^j}{n_i}\log\frac{n_i^j}{n_i}\right) \tag{6}$$

$$Purity=\sum_{i=1}^{k}\frac{1}{N}\max_j(n_i^j) \tag{7}$$

$n_i^j$ is the number of data of $j$th original class assigned to the $i$ th cluster.

## 4.2. Experimental result

Figure 2 shows the changes of $V$ with the

increasing of $l$. And $\beta_{step}$=0.3 and $\beta$=0.9 for NBA data set. For the Zoo data set, $\beta_{step}$=0.3 and $\beta$=1.8.

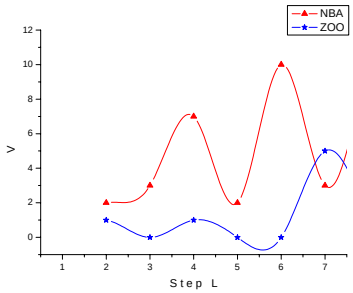


Figure 2 Changes of $V$ with the Increase of Step $L$

Table 2 lists details of detected outliers of the new method. Compared with the algorithm of [5], the new method detects not only the out-performing players according to each single attribute and the all-powerful players, but also the worst player.

Table 2 Outliers Detected by the New Algorithm

| No | Step Num. | Name | *Ppg* | *Rpg* | *Apg* |
|---|---|---|---|---|---|
| 1 | 1 | Dennis Rodman | 4.7 | 15.0 | 2.9 |
| 2 | 2 | Michael Jordan | 28.7 | 5.8 | 3.5 |
| 3 | 2 | Anthony Mason | 12.8 | 10.2 | 4.2 |
| 4 | 3 | Rod Strickland | 17.8 | 5.3 | 10.5 |
| 5 | 3 | Aaron Mckie | 4.1 | 2.9 | 2.2 |
| 6 | 3 | Charles Barkley | 15.2 | 11.7 | 3.2 |
| Average of All data | | | 9.7 | 4.2 | 2.2 |

Table 3 compares the best clustering results of CURED, CURE, ROCK and Frozen. For CURED, the range of $r$ is [3, 5] and $\alpha \in [0.90, 1.00]$. It shows that CURED outperforms the others and obtains much better final clusters than Frozen.

Table 3 Comparison of CURED with Other Algorithms

| ZOO Data Set | | | |
|---|---|---|---|
| | min Entropy | max Purity | $k$ |
| CURED | 0.0592 | 0.8776 | 9 |
| ROCK | 0.0702 | 0.8812 | 9 |
| Frozen | 0.0511 | 0.8515 | 19 |
| Face Feature Data Set ($\beta$=4.0) | | | |
| | min Entropy | max Purity | $\alpha$ / $r$ |
| CURE | 0.377 | 0.466 | 0.97 / 3 |
| CURED | 0.060 | 0.877 | 0.91 / 3 |

# 5. Conclusion

The paper proposes a clustering strategy that utilizes outlier information in the determination of the final cluster number $k$. A dynamic auto-stopped clustering algorithm CURED is stated based on a new outlier-detection algorithm and the improved CURE algorithm. The experimental results show the good performance of CURED in both outlier-detection and clustering.

# Acknowledgements

This work is sponsored by the Natural Science Foundation of China under grant number 60373099 and the Research Foundation of Harbin Engineering University under grant number F2004060.

# References:

[1] C. Rosenberger, K. Chehdi. Unsupervised Clustering Method with Optimal Estimation of the Number of Clusters: Application to Image Segmentation. International Conference on Pattern Recognition, Volume 1. Sep. 2000.

[2] S. Guha, R. Rastogi, K. Shim. CURE: an Efficient Clustering Algorithm for Large Database. In Proceedings of the ACM SIGMOD Conference on Management of Data. Seattle, Washington: ACM Press. 1998. 73-84.

[3] Ana L.N. Fred, José M.N. Leitão. A new Cluster Isolation criterion Based on Dissimilarity Increments. IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 25, No. 8, August 2003: pp944-958.

[4] http://www.ics.uci.edu/~mlearn/

[5] Sridhar Ramaswamy, Rajeev Rastogi, Kyuseok Shim. Efficient Algorithms for mining outliers from Large Data Sets. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. Dallas, Texas, United States, 2000. pp: 427 – 438.

[6] Ying Zhao, George Karypis. Criterion Functions for Document Clustering: Experiment and Analysis. Technical Report #01-40, 2001, University of Minnesota.

[7] Zhang Yue, Zhou Chun-Guang. The Research on Face Recognition Based on the Radial Basis Function Network. Journal of System Simulation, Vol 13 Suppl. pp. 104-107, 2001.