The Discovery Of Attribute Feature Cluster For Any Clustering Result Based On Outlier Detection Technique

Gang Liu[1], Jiming Pang[2], Xiufeng Piao[1], Shaobin Huang[1]
[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, CHINA 150001
[2]College of Computer Science and Technology, Jilin University, Changchun, Jilin, CHINA 130012
{liugang, pangjiming, piaoxiufeng, huangshaobin}@hrbeu.edu.cn

Abstract

The comprehension of the clustering result is a problem that hasn't yet resolve, which having important meaning to the usage of the cluster result and the evaluation of the cluster effect. We put forward the method discovering attribute feature cluster for any clustering result based on outlier detection technique, and put forward an outlier detection algorithm based on even distribution pattern. Through carrying on outlier analysis to all data cluster attribute descriptions, we discovered the feature attribute of each data cluster, and then carried out the comprehension to the clustering result. The remarkable point lies in the method doesn't only aim at a particular clustering algorithm, but also the analysis of any the clustering algorithm result. Experiment to the UCI data set indicated, the method submitted in this paper obtained better result.

Keywords*: attribute feature cluster; clustering; outlier detection; even distribution pattern*

1. Introduction

These As a kind of typical none monitor technique, clustering technology's object is to classify objects into some clusters by their similarity, make the similarity of the data in the same cluster is as possible as small, the similarity of the data of different cluster is as possible as large[1]. At present, people have carried on an extensive research to the clustering method, and appear many clustering methods[2].

But, the comprehension of the clustering result is still a problem that hasn't yet resolve[1]. That problem has important meaning to the usage of the clustering result and the evaluation of the cluster effect. We can mostly comprehend the clustering result from two angles: (1) what is the characteristics of a data cluster, and what is the feature attribute that includes the data cluster? (2) What are the differentiations among some data clusters? At this point, some research adopt visualization technique to display clustering results[3], to aid people to comprehend a clustering result, but face the classic problem of multidimensional data visualization; Another researches give a resolution method to clustering result of a particular algorithm, for example, the FECF algorithm[4] point out the feature attribute of each data cluster.

This paper aims at "the data cluster feature comprehend" problem, and put forward the method discovering attribute feature cluster for any clustering result based on outlier detection technique. Through the covariance of each dimension of data attribute in each data cluster, to acquire the attribute description of each data cluster; and through carry on outlier analysis to all data cluster attribute descriptions, discover the feature attribute of each data cluster, and then carry out the comprehension to the clustering result. The remarkable point is, the method in this paper doesn't only aim at a particular clustering algorithm, but also the analysis of any the clustering algorithm result.

Outlier detection is KDD an independence research domain[1], and its object is to discover the data that have "differential behavior". In principle, outlier detection algorithm is divided into four classes: distribution-based, depth-based, distance-based[5, 6] and density-based[7]. The first two methods exist in some fatal shortage, for example, it isn't suitable to handle multidimensional data, and need to suppose the data distribution method in advance. Therefore, the research emphasis is distance-based outlier detection algorithm, and put forward an outlier detection algorithm based on even distribution pattern, used to analyze attribute feature cluster in clustering result.

The full paper contents arrangement is as follows: section 2 introduces an outlier detection algorithm based on even distribution pattern, section 3 introduces the method discovering attribute feature cluster for clustering result based on outlier detection technique,

978-0-7695-3112-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICICSE.2008.37

section 4 analyzes the experiment, in the end summary full paper.

2. Outlier Detection Algorithm Based on Even Distribution Pattern

The basic idea of distance-based method is: if the distances between data a and most other data are larger than the threshold D_{out}, a is an outlier. However, this method ignores the local distribution feature of one data. For instance, in Figure 1, p is the first outlier candidate because it is the farthest one from the others, but data locating near q are more compressed than those near p, thus q is a more natural outlier. And it is critical for a distance-based outlier mining algorithm to decide an appropriate threshold D_{out}.

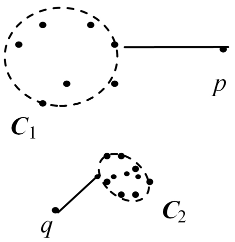


Figure 1. Local Distribution Situation of Data

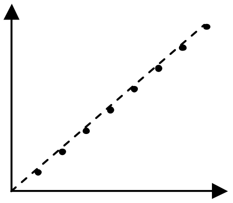


Figure 2. The case with $D_{out} < \overline{D}_{NN}$

To overcome these shortcomings, the paper proposes an outlier detection algorithm which refers to data's even distribution pattern and takes data's local distribution feature into consideration.

The even distribution pattern is a useful reference, for outliers exist when data distribute unevenly. Obviously the distances $\overline{D}_{NN}$ between each data and its nearest neighbor are the same in this case. To compute $\overline{D}_{NN}$, the M-dimensional vector space S that N data locate in is equally divided into N grids with only one data in grid's center. Easy to prove that the length of ith-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 ith-dimension. And the super-rectangle R decided by $a_{\max}$ and $a_{\min}$ contains S. Thus:

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

To mine outliers under the realistic distribution situation, two factors are taken into consideration.

First, parameter β is adopted to describe the diversity degree of the realistic distribution from the even distribution pattern and $D_{\text{out}}=\overline{D}_{NN}/\beta$. The factors that influence the value of β are: (1) when data cluster together, the average distance between data or clusters will be larger than $\overline{D}_{NN}$, which means $D_{out} > \overline{D}_{NN}$, like the case of Figure 1; (2) the space R decided by $a_{\max}$ and $a_{\min}$ is larger than the space S, for example data are distributed on a line like Figure 2, in that case $D_{out} < \overline{D}_{NN}$.

The following method is proposed to decide the value of β:

(1) suppose β=β_{Step} when only an outlier is detected and call β_{Step} as the *step length*, easy to prove that $\beta_{Step} = D_{NN}(a_{farest}) \times \xi(a_{farest}) / \overline{D}_{NN}$ and a_{farest} satisfies $D_{NN}(a_{farest}) \times \xi(a_{farest}) \geq D_{NN}(b) \times \xi(b)$ for any b;

(2) observe the increasing speed V of the detected outlier number n_{out} under different value of β, viz. $V = \nabla n_{out} / \nabla\beta = \nabla n_{out} / \beta_{Step}$, where $\beta = l \times \beta_{Step}$ and l={1, 2 …}, call l the *step Num.*;

(3) if V reaches its first peak when $l_i \times \beta_{Step}$, $\beta = (l_i - 1) \times \beta_{Step}$.

This method is based on following observation: n_{out} increases much faster with further decreasing of D_{out} after outliers are detected. That is because outliers are extremely far away from the others while the normal are relatively near to each other. Taking Figure 1 for an example, after p and q are recognized as outliers, many data in cluster C_1 and C_2 will be detected as outliers with the increase of *step num. l.* Figure 3 shows the detail.

Second, factor ξ 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 $D_{\text{NN}}(b)$. The value of $\xi(a)$ shows the isolation degree of a from its neighbors. For example, $\xi(q) > \xi(p)$ in Figure 1. The special method should be applied for the very similar or duplicate data. The reason is that: suppose that b is a's nearest neighbor and b and c are very similar or the

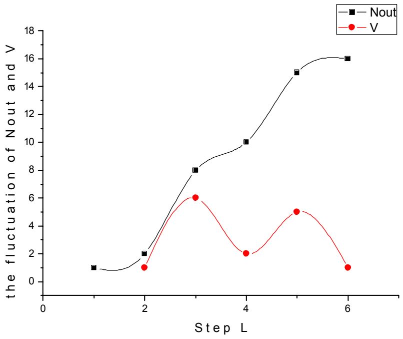


Figure 3. Example of the Changes of n_{out} and V

same, which means $D_{NN}(b)\rightarrow 0$, Viz. $\xi(a)\rightarrow\infty$; thus a would be regarded as an outlier no matter what value $D_{NN}(a)$ is. To avoid this, adjust the equation for $\xi(a)$ as follows:

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

Therefore, the outlier evaluation criterion is stated as follows:

Data a is an outlier, if

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

The new outlier-mining algorithm is constructed on the proposed outlier evaluation criterion. Its overview of is listed in Figure 4. Since $a_{\max}$ and $a_{\min}$ can be decided in data reading, they are the input of the algorithm.

```
Void outlier($a_{\max}$, $a_{\min}$)
{   decide the value of $\beta$;
    $D_{out}=\sqrt{\sum_{i=1}^{M}((a_{\max}^{(i)}-a_{\min}^{(i)})/\sqrt[M]{N})^{2}}/\beta$;
    for ($i$=0; $i$<$N$; $i$++)  //evaluating the $i$th data
    { call the nearest neighbor of data $i$ as data $j$ ;
        if ($D_{NN}(j)$>=$10^{-4}$)
            $\xi(i)=D_{NN}(i)/D_{NN}(j)$;
        else  $\xi(i)$=1;
        if ( $D_{NN}(i)*\xi(i)>D_{out}$ )
            data $i$ is an outlier;
    }
}
```

Figure 4. The New Outlier Detection Algorithm

3. Discovering Attribute Feature Cluster Based on Outlier Detection Technique

For discovering the corresponding feature cluster of attribute m_i, it is needed to analyze in arbitrarily of the data cluster C_j to bunching circumstance of attribute m_i in the clustering result. According to the attribute m_i's span from minimum value to the maximum value [min_i, max_i], we can divide the attribute m_i into some sub-spans, statistics all the number of attribute m_i value located in each sub-span in data cluster. This statistical result can be used to the distribution circumstance of the data cluster C_j to the attribute m_i. Through analyzing the distribution circumstance of all the data clusters, we can discover the data cluster which distribution circumstance is especially differential, as the feature cluster of the attribute m_i.

The detailed analysis procedure is as follows:

1. Record the result of arbitrarily clustering algorithm like $\{C_1, C_2\ldots\ldots C_n\}$, as the input;

2. To the attribute m_i, the minimum and maximum value span [min_i, max_i] of attribute m_i of all data, is divided into l spans;

3. To the data cluster C_j, might as well establish the total data amount is n_j among them, statistics all the number of attribute m_i value located in each sub-span in data cluster, note as $\{n^1_j, n^2_j\ldots\ldots n^l_j\}$; distribution circumstance of the data cluster C_j to the attribute m_i is noted as $\{n^1_j/n_j, n^2_j/n_j\ldots\ldots n^l_j/n_j\}$;

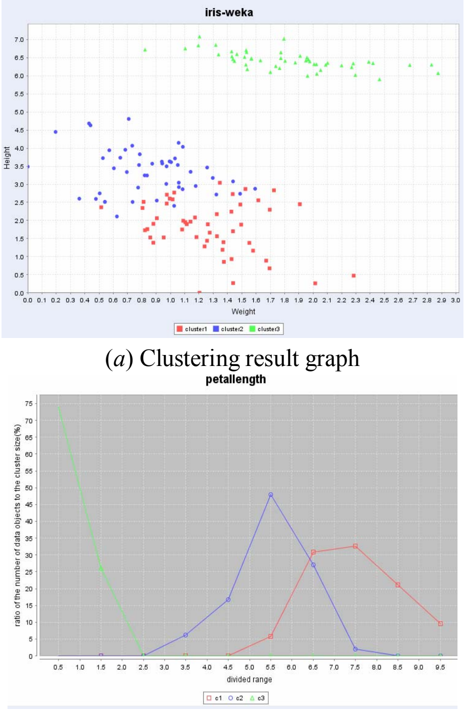
4. Repeat step 3, get the distribution circumstance of all data clusters for the attribute m_i, adopt the detection algorithm in section 2 to discover the outlier, the corresponding cluster of this outlier is just the feature cluster of the attribute m_i;

5. Repeat step 2 to 4, obtain the feature clusters of all attributes.

4. The Experiment and Analysis

In this paper, we use the clustering result of UCI's iris, ZOO and Housing data set to test the analytical result of the method to the attribute feature clusters originally.

Figure 5. (a) shows the visualization result to adopt FastMap algorithm[9] to each data cluster in Iris data set, among them, the clustering result is the result setting the result data cluster to [3-5] and using X-Means algorithm[2]. When adopting the algorithm mentioned in section 3, we analyze and notice, the data cluster C_3 is the feature cluster of all attributes. For illuminating analysis result clearly, Figure 5. (b) shows the distribution circumstance of 3 result data clusters to the attribute petallength, we can obviously see the distribution abnormity between data cluster C_3 and other two data clusters. The analytical result to the other 3 attributes is similar to this. This result can also validated by Figure 5. (*a*), we can see whatever attribute is adopt, the data cluster C_1 and C_2 always cross distribution, and it is very difficult to distinguish each other.

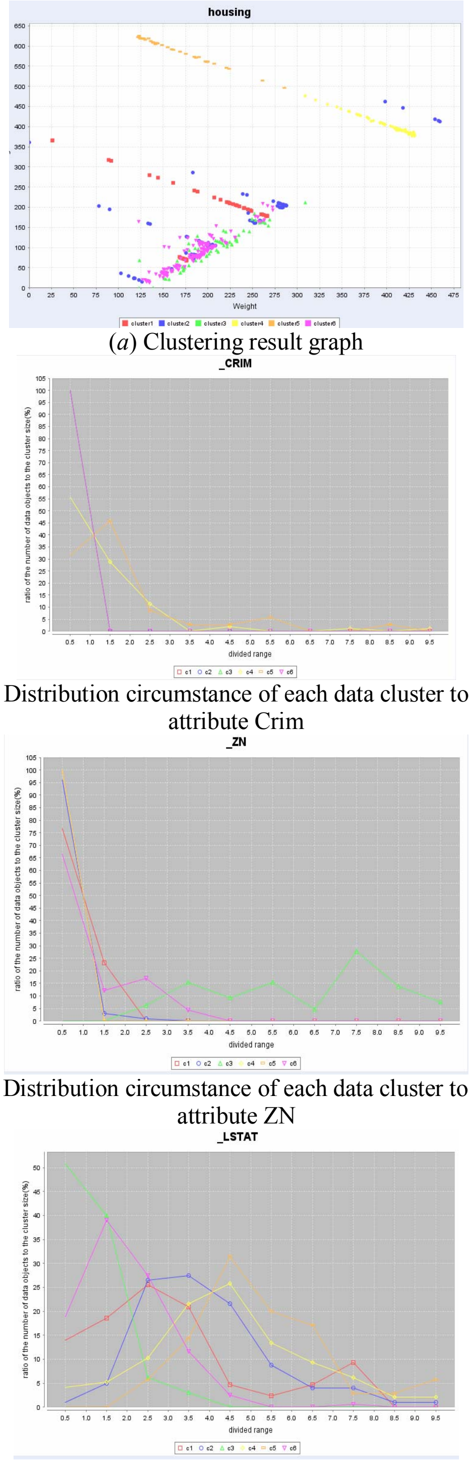


(*a*) Clustering result graph

(*b*) Distribution circumstance of each data cluster to the attribute petallength

Figure 5. The clustering result attribute feature cluster of the Iris data set

Figure 6. (*a*) shows the visualization result to adopt FastMap algorithm[9] to each data cluster in Housing data set, among them, the clustering result is the result using Frozen algorithm[10]. When adopting the algorithm mentioned in section 3, we analyze and notice, the data cluster C_3 is the feature cluster of ZN and LSTAT attributes. For illuminating analysis result clearly, Figure 6. (*c*) to (*d*) show the distribution circumstance of each result data clusters to the attribute ZN and LSTAT, we can obviously see the distribution abnormity between data cluster C_3 and other two data clusters. Moreover, as Figure 6.(*b*) shows, to the attribute CRIM, we can discover the distribution curve of data cluster C_1,C_2,C_3,C_6 is consistent, so it is distinguished with the data cluster C_4 and C_5, and becomes the attribute feature cluster of the attribute CRIM.



(*a*) Clustering result graph

(*b*) Distribution circumstance of each data cluster to the attribute Crim

(*c*) Distribution circumstance of each data cluster to the attribute ZN

(*d*) Distribution circumstance of each data cluster to the attribute petallength

Figure 6. The clustering result attribute feature cluster of the Housing data set

Table 1 shows a data cluster that adopts DBSCAN algorithm[11] to analyze ZOO data set and adopts the analytical result of section 3 algorithm. During analyzing, we find, in the data cluster C_4, although various animal is all kinds of birds and beasts, but generally, all of them have no tail, so the feature cluster of the attribute tail behave better. But in data cluster C_3, generally, various animals are aquicolous predatory fish, so the feature clusters of the attribute Fins and Predator behave better.

Table 1. Detailed circumstance of the clustering result attribute feature cluster of ZOO data set

Attribute	Cluster number	Detailed circumstance inside the cluster
tail	C_4	Wasp, termite, starfish, worm, lobster, ladybird, octopus, moth, flea, gnat, housefly, honeybee, slug, clam, seawasp, crayfish, crab
Fins, Predator	C_3	sole, haddock, dogfish, piranha, chub, bass, tuna, pike, carp, herring, catfish

5. Discussion

This paper put forward the method discovering attribute feature cluster for any clustering result based on outlier detection technique, and put forward an outlier detection algorithm based on even distribution pattern. The remarkable point is, the method in this paper doesn't only aim at a particular clustering algorithm, but also the analysis of any the clustering algorithm result. Experiment to the UCI data set indicates, the method submitted in this paper obtained better result. The next work moves to comprehend a clustering result from other angles, including understanding the reason to result in the discrepancy between arbitrarily two data clusters.

References

[1] Jiawei Han, *Data Mining: Concept and Technology*, Mechanism Industrial Publishing Company, 2001

[2] Dan Pelleg, Andrew Moore, “X-means: Extending K-means with Efficient Estimation of the Number of Clusters,” *Proc. 2000 Int. Conf. on Data Mining*, 2000, pp. 727-734

[3] T.C. Sprenger; R. Brunella; M.H. H-Blob Gross, “A Hierarchical Visual Clustering Method Using Implicit Surfaces,” *Visualization 2000, Proceedings*, 2000, pp. 61-68

[4] Tomotake Nakmura, “Feature extraction of clusters based on flexdice,” *Proceedings of the 21st International Conference on Data Engineering Workshops*, 2005, pp. 1126-1130

[5] M. Edwin. E. Knorr, R. Ng, “Finding intentional knowledge of distance-based outliers,” *Atkinson MP, Orlowska ME, Valduriez P, Zdonik SB, Brodie ML, eds. Proc. of the VLDB Conf. Edinburgh,* Morgan Kaufmann Publishers, 1999, pp. 211-222

[6] S. Ramaswamy, R. Rastogi, K. 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

[7] M. M. Breunig, H. P Kriegel., R. T. Ng, et al. “LOF: Identifying Density-based Local Outliers,” Chen, W., Naughton, J.F., Bernstein, P.A., eds. *Proceedings of the ACM SIGMOD International Conference on Management of Data,* Dallas, Texas: ACM Press, 2000, pp. 93-104

[8] S. Hettich &, C.L. Blake &, C.J. Merz, “UCI Repository of Machine Learning Databases.” *Irvine, CA: University of California, Department of Information and Computer Science,* 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html

[9] C. Faloutsos, K. Lin, “FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets,” *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data,* San Jose, California: ACM Press. 1995, pp. 163-174

[10] 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, 2003, pp. 944-958

[11] M. Ester, H. P. Kriegel, J. Sander, X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, OR, 1996, pp. 226-231