Hierarchical Clustering Algorithm Based on Granularity

Jiuzhen Liang[1,2], Guangbin Li[1]
[1]Zhejiang Normal University, Jinhua, 321004, China
[2]Center for Machine Perception, Faculty of Electrical Engineering,
Czech Technical University, Technicka 2, 16627 Praha 6, Czech Republic
liangjz@zjnu.cn, liang@cmp.felk.cvut.cz, wesley2005@126.com

Abstract

This paper proposes a hierarchical clustering algorithm based on information granularity, which regards clustering on sample data as the procedure of granule merging. In the promoted algorithm, firstly each sample is named with an initial class, then for a given granular threshold those pairs of samples, whose distance among them is less than the threshold, will be merged to one class and generate a new larger granule. Repeat this procedure until certain conditions are satisfied. This paper also discusses computational complexity of the novel algorithm and compares them with the traditional hierarchical clustering algorithm. In the last, some experimental examples are given, and the experimental results show that this algorithm can efficiently improve the clustering speed without affecting the precision.

1. Introduction

As the old saying goes "Birds of a feather flock together", clustering is an important style of people's recognition. Clustering, according to certain attributes of object, divides a sample set into some different classes. It makes the similarity among classes as small as possible, and the one in the class as big as possible. The clustering technique is widely used in many areas, such as data statistics, pattern recognition, machine learning and data mining [1].

Hierarchical clustering is a commonly used algorithm. There are many types of classical hierarchical clustering algorithms, such as BIRCH [2], CURE [3], ROCK [4], and Chameleon [5]. Although most of these algorithms have powerful abilities in clustering, limited to their high computational complexity, they do not adapt dealing with large scale of sample sets.

This paper proposes a hierarchical clustering algorithm based on information granularity, which utilizes hierarchical clustering to sample data based on different granularity. Experiments show that this algorithm can effectively improve the rate of hierarchical clustering without influence on the clustering results.

This paper firstly introduces some basic knowledge on hierarchical clustering, agglomerative clustering and information granularity; then describes the hierarchical clustering algorithm based on granularity(HCAG) in detail; also analyzes the algorithm computational complexity; lastly, gives some experimental results.

2 Hierarchical clustering algorithm(HCA)

In hierarchical clustering, data is decomposed hierarchically step-by-step. According to the procedure of generating hierarchical structure, there are two types of hierarchical methods: agglomeration and partition. If data is decomposed hierarchically by manners of bottom-up, it is called agglomerative hierarchical clustering. If it is by manners of top-down, it is called divisive hierarchical clustering.

In agglomerative hierarchical clustering, firstly each sample is labeled with a class, then gradually those samples in different classes are merged into some larger ones, until all the samples are in the same class, or satisfy certain terminal condition. On the other way round, in divisive hierarchical clustering first all the samples are put in one class, then the class is gradually divided into smaller and smaller subclasses, until each sample becomes one class, or satisfy certain terminal condition, such as getting some desired number of classes, or the distance between the nearest two classes is smaller than a given threshold.

In fact, agglomerative clustering corresponds to constructing certain data structure of a tree. Its leaves stand for N samples; the root is the last one when all samples are merged into one class, and those levels in the middle stand for states of data during the procedure of clustering.

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.53

3 Information granularity and clustering

Granularity originally refers to "the average measure that is in the same size as atoms". It is used as "the average measure of information granularity". Information granularity is the measure of the reification to information and knowledge in different levels [6].

B. Zhang and L. Zhang proposed the concept of information granularity in 1990, and gave some thorough incisive comments [7] as the following. Artificial intelligence and cognitive science researcher have also observed that one common characteristic of artificial intelligence is that when perceiving and processing problems in the real world, it usually utilizes different strategies in different levels, and also observes and analyzes the same problem in different granules. The concept of information granularity contributes to formalizing such cognitive ability for human being.

3.1 The relationship between different granules

For certain problems, sometimes it needs solutions in both big and small granules. So it is necessary to study the relationship between different granules.

Suppose R is the set of all equivalent relation in universe U. The equivalent relation, in other words, the "big" and "small" of granule, can be defined as follows.

Theorem 1 $R_1, R_2 \in R$, if for any $x, y \in X, xR_2y \rightarrow xR_1y$, it is said that R_2 is smaller than R_1, namely $R_2 \leq R_1$.

The following theorem can be proved referring to[7].

Theorem 2 R forms a completed partial order grid in the mean of relationship "$\leq$" defined in theorem 1.

This is a very profound theorem, because it reveals the essential property of granularity, and sets up an important foundation for other properties. The proof for the detail refers to [7].

Based on theorem 2, the following sequence is true clearly.

$$R_n \leq R_{n-1} \leq \cdots \leq R_1 \leq R_0 \tag{1}$$

Intuitively, sequence (1) can be mapped to a tree T with n levels. All of its leaves can form a set X. The nodes in each level match up to one partition. The hierarchical diagram corresponding to the procedure of clustering appears just as a tree with n levels. So, it is sure that there is one equivalent relation matching to it. This is the reason why clustering closely links to granularity.

3.2 The principle of granularity in clustering

Essentially, the results of clustering is the relation of equivalence defined among samples. Two samples, which belong to the same class, are considered equivalent in the measure of certain threshold. We can say that there are some close features between them. A relation of equivalence matches a partition in a sample set, and it divides the set into some subsets, each of which is corresponding to a class generated by clustering.

Along with a series of threshold varying from big to small, it will generate a series of equivalent relation. When using the small threshold, we can describe the slight degree of difference among samples. When using the big one, the rough outline of sample set appears, while some minor details are neglected.

4 Hierarchical clustering algorithm based on granularity(HCAG)

4.1 The idea and basic steps of HCAG

For the sake of discussing clearly, we take the process of agglomeration as the one that the granule changes from small to big. In the process of agglomerative clustering, we only find out the nearest two classes and merge them at one time. As a result, the change of granule is very imperceptible. We can get a relatively good clustering result in this way. But, the granule is slowly getting bigger. So, the computational complexity is rather high. And it cannot deal with clustering sample problem with larger scale.

In this paper, we propose a hierarchical clustering algorithm based on granularity (HCAG), which is also one of agglomerative algorithms. The basic idea is as follows. At first, the original samples form a class. Then, for a given distance threshold θ, we merge all samples(or subclasses) among which the distance threshold is smaller than θ. Lastly, this process will be repeated until all the subclasses are merged into one or iteration satisfies other given stop conditions.

We discuss the problem in detail as follows. Given the initial sample set $X = \{x_1^{(0)}, x_2^{(0)}, \cdots, x_{N_0}^{(0)}\}$ with dimension n. The weight vector corresponding to all the samples is $W = (w_1^{(0)}, w_2^{(0)}, \cdots, w_{N_0}^{(0)})$, where $w_i^{(0)}$ stands for the initial weight of sample $x_i^{(0)}$. Denote K as the desired class number for clustering.

In the k-th loop of mergence, where $k \geq 1$, the centroid of samples in group j can be computed as follows,

$$x_j^{(k)} = \frac{\sum\limits_{x_i^{(k-1)} \in C_j^{(k-1)}} w_i^{(k-1)} x_i^{(k-1)}}{\sum\limits_{x_i^{(k-1)} \in C_j^{(k-1)}} w_i^{(k-1)}} \tag{2}$$

where $i = 1, 2, \cdots, N_{k-1}$ and $j = 1, 2, \cdots, N_k$. And after mergence, the weight of new sample(subclass j) is

$$w_j^{(k)} = \sum_{x_i^{(k-1)} \in C_j^{(k-1)}} w_i^{(k-1)} \tag{3}$$

While before mergence, the distance between two samples, namely p and q, is

$$d_{pq}^{(k-1)} = \|x_p^{(k-1)} - x_q^{(k-1)}\| \tag{4}$$

Here we describe the Hierarchical Clustering Algorithm based on Granularity(HCAG) as follows:

Step 1. Initialize parameter α, iteration number record $k = 1$ and the desired number of clustering K;

Step 2. Label each sample as a single class, i.e. $C = N$;

Step 3. Run the following sub-steps for all samples:

Step 3.1. Compute the distance among classes $d_{pq}(1 \leq p, q \leq N)$, the maximum and the minimum distance $d_{\max}, d_{\min}$;

Step 3.2. Compute the granular threshold θ according to the formula:

$$\theta_k = (1 - \alpha)d_{\max} + \alpha d_{\min} \tag{5}$$

Step 3.3. Any two classes p and q, merge into a new one if their distance d_{pq} is less than θ_k. After all classes merge under the θ_k threshold level, compute the new class coordinate according to equation (2).

Step 4. If $C \leq K$, end; else $k = k + 1$, goto step 3.

In HCA, only the nearest two classes merge each time. While in HCAG, all the classes, whose distance are less than θ, will merge. As a result, one more class merge each time. So, comparing to HCA, HCAG evidently improves the clustering speed.

4.2 Algorithm analysis

In HCA, each time only the nearest two samples merge. So, N samples finally merge into one class, and it needs $N - 1$ iterations. Considering the k-th iteration, the current number of classes is $N - k + 1$. In order to compute the distance among the $N - k$ classes, it needs

$$C_{N-k+1}^2 = \frac{1}{2}(N - k + 1)(N - k) \tag{6}$$

So, the total computational complexity for HCA is

$$\Omega_{HCA} = \frac{1}{2}\sum_{k=1}^{N-1}(N - k + 1)(N - k) = \frac{1}{6}(N^3 - N) \tag{7}$$

While in HCAG, if we merge $S + 1$ samples each time, i.e., after mergence the total number of classes reduces by S, the total iteration is N/S. Considering the k-th iteration, the current number of classes is $N - (k - 1)S$. Computing the distances among the $N - (k - 1)S$ samples, it needs

$$C_{N-(k-1)S}^2 = \frac{1}{2}(N - (k - 1)S)(N - (k - 1)S - 1) \tag{8}$$

So, the total computational complexity of HCAG appears to be

$$\begin{aligned}\Omega_{HCAG} &= \frac{1}{2}\sum_{k=1}^{N/S}(N - (k - 1)S)(N - (k - 1)S) \\ &= \frac{N^3}{6S^3} + \frac{(1 - S)N^2}{4S^2} + \frac{(1 - 3S)N}{12S}\end{aligned} \tag{9}$$

From equation (7) and (9), we have the following conclusions. If $S = 1$, i.e. $\Omega_{HCAG} = \Omega_{HCA}$, which means HCAG turns out to be HCA. While if $S > 1, \Omega_{HCAG} << \Omega_{HCA}$. Clearly, compared to HCA, HCAG can sharply shorten the clustering time. But, the number of mergence S cannot be too large, or it will directly influence the clustering precision.

In HCAG, the option of threshold θ will directly influence clustering algorithm. The two classes, whose distance is less than θ and seemed as neighbors in the current granularity, will be merged each time in the process of iteration. If θ is too big, too many classes will be merged. Even it is possible that a large mount of classes is merged into one class at one time, which leads to loss of clustering accuracy. On the contrary, if θ is too small, very few classes are merged each time. This will slow down the speed of clustering.

Firstly, θ must be more than the minimum distance between current classes namely $d_{\min}$, otherwise there will be no samples to merge. Secondly, θ must be between $d_{\max}$ and $d_{\min}$ and needs to be a little larger than $d_{\min}$. In this way, we can not only improve the rate of clustering convergence, but also guarantee the clustering precision. Therefore, we choose θ by the equation (5). In addition, θ depends on the distributing of distance among classes. Figure 1 shows the relation between distance threshold and the number of class pairs(or subclasses) in sample set(S4) as in the following experiment.

In Figure 1, the abscissa denotes the value of distance threshold θ. The dots on the curve stand for numbers of class pairs whose distance is less than the current threshold θ. For example, the current number of samples is 250, if we do not expect to have the mergence number more than $250 \times 10\%$, we need to find out the abscissa of which the number of class pairs is 25. By counting experimental data, the current θ is 4.3502×10^{-4}. Matching it, 5×10^{-4} is a suitable value for α in equation (5), thereby giving guide to the choice of parametric θ or α in HCAG.

We can also choose the merging samples in another way. Find out the smallest distance among the former S samples and merge these classes together. But, it spends some time on computing the smallest distance.

5 Experiments

In this experiment, four groups of 2 dimensions data S1, S2, S3 and S4 are tested for clustering, using HCA and

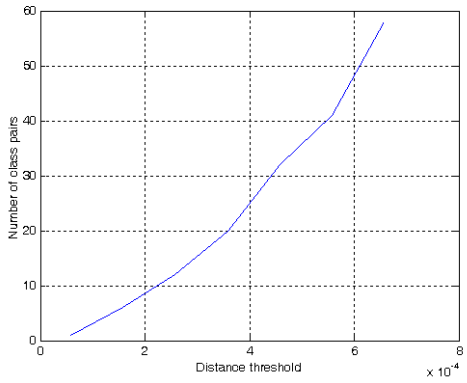

Figure 1. The relation between distance threshold and number of class pairs

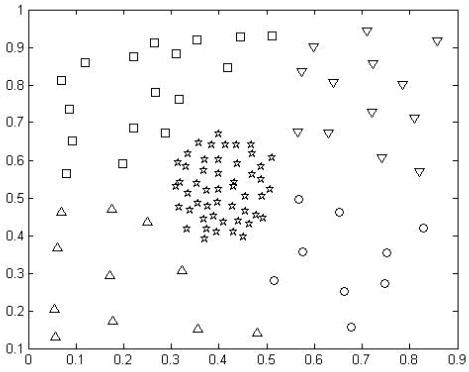

Figure 2. The distribution of sample set S4

HCAG respectively. The sample set S4 distribution is in figure 2. In HCAG, we choose $\alpha = 0.0005$. The experimental results for the two algorithms are showed as in Table 1.

Table 1. Comparison of HCA and HCAG

Data	N	K	HCA(i)	HCA(s)	HCAG(i)	HCAG(s)
S1	100	5	95	10.969	34	1.515
S2	100	3	97	10.469	30	1.906
S3	100	3	97	10.328	34	1.687
S4	250	5	245	113.08	35	6.641

In table 1, column HCA(i) denotes HCA iteration number, while HCA(s) is the runtime of HCA. The iteration of HCA algorithm is $N - K$, and in HCAG the iteration is smaller than the former. Because HCAG combines more than one class each time, it causes rapid reduction of classes. The runtime of the HCAG is one fifth or even less than HCA's in the four sample sets, especially for sample set S4. Because HCA combines two samples each time and there exists a great deal of small classes in the initial time of clustering, it needs much more time to compute the distance between two classes.

6 Conclusions

Referring to the idea that information is clustered on basis of different levels in information granularity, and taking the traditional process of agglomerative methods as granule changing from small to big, this paper proposes a novel hierarchical clustering algorithm. Comparing to the traditional algorithm, its computational complexity is relative lower and easy to implement. Experiments show the efficiency of the novel algorithm. In addition, for different problems, the capability of HCAG depends on the granule threshold θ. This paper proposes some pieces of instructive suggestions. It still needs more research on the problem how to choose more reasonable θ.

Acknowledgement This work is partly supported by CMP laboratory, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University. The first author is supported by the Exchange Project Agreement between the Czech Ministry of Education and the Chinese Ministry of Education.

References

[1] A. K. Jain, M. N. Murty, P. J. Flynn. Data clustering: a review. ACM Computing Surveys, 1999, 31(3): 264-323.

[2] T. Zhang, R. Ramakrishnan, M. Livny. An efficient data clustering method for very large databases. Magement of data, 1996, 103-114

[3] R. Rastogi,S. Guha, K. Shim. CURE: an efficient clustering algorithm for large database. Information Systems, 2001, 26(1): 35-58

[4] S. Guha, R. Rastogi, K. Shim. A robust clustering algorithm for categorical attributes. Data Engineerin, 1999, 512-521

[5] G. Karypis, E. H. Han, V. Kumar. A hierarchical clustering algorithm using dynamic modeling. COMPUTER, 1999, 32(8): 68-75

[6] J. Shao. Information granularity computing based on rough sets. Institute of Automatics, Chinese Academy of Sciences, Beijing, 2000.

[7] B. Zhang, L. Zhang. Theory of Problem Solving and Its Application. Beijing: Tsinghua University Press, 1990