2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)
IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V51.59

Clustering Algorithm for Mixed Attributes Data Based on Restricted Particle Swarm Optimization

Zhang Tiejun[a,b+], Yang Jing[a], Zhang Jianpei [a]

[a]Department of Computer Science, Harbin Engineering University, Harbin,150001, China
[b] Department of Computer Science, Northeast Agricultural University, Harbin, 150030, China

Abstract. When the data mixed with numerical and categorical values is processed at present, it is very common to convert the categorical values into numerical ones and then cluster them according to a certain weight. Obviously such clustering results rely heavily on the weight given by experts. Hence in this paper a categorical attribute is proposed as a potential field restriction to limit the searching directions of particle swarm, which consequently improves the speed and effectiveness of the clustering algorithms.

Keywords: numerical and categorical values, potential field restriction, particle swarm

1. Introduction

Clustering is a process of classifying many individual objects into groups of similar ones according to some of their attributes. Its target is to make the objects within the cluster as similar as possible as well as to make the objects between the cluster as different as possible. Most of the present clustering algorithms are used to process the data with a single attribute, such as numerical values or categorical values. However, in the clustering practice, what needs processing is not only the data with numerical values but also the mixed one with categorical values including texts and images. Clearly these algorithms, when they are used to process data with mixed attributes, should all be correspondingly converted – totally converted into numerical values or categorical ones to be processed. The precision of the clustering is affected when converted. What distinguishes clustering from classifying is that the former isn't guided by priori or background knowledge, but an auto-detecting unsupervised learning process, which is based on the object similarities. But the clustering users usually have a clear demand for its application. In view of this, in many of its application, the most effective way is to be more inclined to add the users' preferences or restrictions into the clustering process in order to greatly affect the results of acquiring knowledge. Thus it can help us find a knowledge pattern which interests more users and meets more users' needs. As for the clustering with categorical attributes, Ganti V, Guha S and Andritsos P proposed different processing methods respectively. Ganti V proposed CACTUS[1] algorithm on the basis of clustering the summary information of categorized data sets. Guha S proposed ROCK[2] by measuring the similarity between records with linking numbers between two records. Such algorithm is hierarchical clustering algorithm, whose time complexity is the cube of the number of the data. It is not applicable to process large data sets. Andritsos P made use of the nearest information theory which is based on an information bottlenecked framework and proposed LIMBO[3]. It is also hierarchical clustering algorithm with more flexibility. But in view of the past clustering with categorical attributes being too much dependent on distance matrix, Barbara D proposed COOLCAT[4] algorithm which is based on less entropy of similar cluster compared with that of different one. Darshit Parmar, having considered the problem of uncertainty in the clustering process, proposed MMR[5] with the idea of fuzzy set theory and achieved the clustering of data sets with categorical values.

+ Corresponding author. Tel.: +86-451-55191036; fax: +86-451-55190672..
E-mail address: jun.zh_t@163.com.

For the data with numerical values, Agrawal R proposed CLIQUE[6] algorithm which discretes the successive data by means of the orderly intensive grids. And other clustering algorithms such as DOC[7], PROCLUS[8], ORCLUS[9] all need calculating numerical values' statistic.

However, in reality, the above algorithms can't be used in the large number of mixed data with numerical and categorical values. In this case, Huang put forward k-modes[10] algorithm and k-prototypes[11] algorithm and clustered the data mixed with numerical and categorical values. To better solve the problem of the mixed attributes clustering, my paper proposes Restricted PSO Clustering (RPSOC). It first clusters data with categorical values, takes every cluster as a restricted source to form a restricted grads field, and then guides particle swarm to preferably search in the restricted grads field, and finally finds optimization result and forms clustering in the feasible domain.

2. Algorithm foundation

2.1. Potential field model

If the data sample is composed of a finite particle and per unit consist of d-typed substance of different ingredients, the gravitation model concept can be introduced to the clustering process. The detailed illustration is as the follows.

The known space $\Omega \subset R^d$ includes n numbers of objects of data sets $D = \{x_1, x_2, \cdots, x_n\}$, the related concept of the RPSOC algorithm can be illustrated as follows:

The gravitation concept model:

Gravitation function:

The gravitation $f_{x \in C_j}(A_i, C_j)$ received by particle $x \in \Omega$ is the restriction force of A_i on cluster C_j, the Equation as (1)

$$f(A_i, C_j) = G \frac{M(A_i) M(C_j)}{R^2(A_i, C_j)} \tag{1}$$

In the above equation, M is the number of particles in various cluster. And R is Euclidean distance between the center of restricted source A_i and the center of cluster C_j.

Global gravitation function: it can be defined as the sum of all the restriction force of cluster C_j received from A, for set $A = \{A_1, A_2, \cdots, A_m\}$ has m centers of potential field, the definition of global gravitation function as (2):

$$f_{x \in Cj}(C_j) = \sum_{i=1}^{m} f_{x \in C_j}(A_i, C_j) \tag{2}$$

Restriction of potential field: the center of the field A is known, and if the subset $C \subset D$ exists, $\forall x \in C$ and x restricted by potential field A, and $f_{x \in C}(A) \geq \xi$ (ξ is preset threshold), C is restricted by potential field A. The direction of the restriction is the one of gradient of the potential field as well as the one for velocity potential restriction.

Acceleration:

$$a_j = \frac{f(C_j)}{M(C_j)} = G \sum_{i=1}^{m} \frac{M(A_i)}{R^2(A_i, C_j)} \tag{3}$$

Gravitation coefficient

G=1 can be omitted.

The introduction of potential field restriction linked with the bond of space information better removes the disadvantage of separation between numerical and categorical attributes and provides practical problem-solving method for their mutual coupling. By means of non-numeric value attribute information of space objects, equipotential line can be formed. According to the distribution features of the equipotential line, the tendency for the some part of data to cluster in space can be found out. The changes in restriction potential values show the intensity of interaction of spatial objects, which as a reciprocal restriction reflects the neighboring information and distribution features of the spatial objects. As is shown in Fig.1:

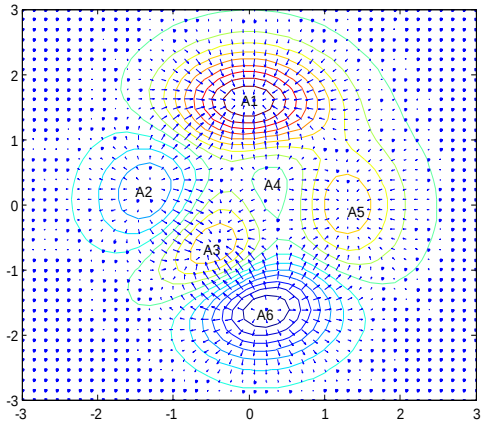

Fig.1.Restriction potential gradient

The restriction source $A=\{A_1,A_2,\cdots,A_6\}$ forms gradient restriction in space, and the restriction direction is shown in it.

2.2. Particle swarm optimization

Based on swarm intelligence PSO is an evolutionary calculating technique proposed by Kennedy and Eberhart. Its advantages are easy realization, high precision and fast convergence and it has already exhibited its advantages in solving the practical problems.

In PSO algorithm, particle swarm searches in a d-dimensional space. If the group scale of particle is n, the place for the ith particle can be shown as Xid, Which stands for candidate solution of the problem in the searching space. The degree of solution from good to bad is determined by fitness function F. In every iteration, particle renews it place by tracing the two extremum. individual extremum expressed in Pid, global extemum expressed in Pgd in global PSO, local extemum expressed in Pld in local PSO. Every particle has its own speed known as Vid and it renews it speed and place at the time of t+1 according to the equation followed:

$$V_{id}(t+1)=\omega V_{id}(t)+c_1 rand()(P_{id}(t)-X_{id}(t))+c_2 rand()(P_{gd}(t)-X_{id}(t)) \quad (4)$$

$$X_{id}(t+1)=X_{id}(t)+V_{id}(t) \quad (5)$$

Vid shows the speed of the ith particle at the dth-dimension. inertia weight ω is scaling factor related with the last speed. c1, c2 is constant as learning factor or accelerating coefficient, rand() is random number distributing uniformly in (0,1).

3. RPSOC

3.1. Improvement

RPSOC mainly improves the standard PSO in the three following way:

- According to the categorical values of the clustered objects, the algorithm adopts LIMBO to create the restriction potential field. The acceleration restriction is formed in space, that is, speed gradient. The initial speed V_{id} in the particle swarm adopts the acceleration value of the present position.
- Because the algorithm restricts the global researching process, accelerating learning factor is no longer needed. The simplified equation is:

$$V_{id}(t+1)=V_{id}(t)+rand()(P_{id}(t)-X_{id}(t))+rand()(P_{gd}(t)-X_{id}(t)) \quad (6)$$

  In the searching process, if a particles moves to a position the mark of which belongs to some cluster, its fitness function value is 0. In this way, the repeated clustering of the clustered data can be avoided.
- When searching within the neighborhood of the present space, the searching direction is non-negative speed gradient, that is, non-decrease direction of acceleration. If fitness function value of the object and its neighborhood is less than preset value, we marked it with cluster number.

3.2. the basic steps of RPSOC

On the basis of the above illustration, the basic steps of RPSOC are as follows:

(1) Create the restriction source: For the data with one categorical attribute, breadth-first search method can be used. But for the data with many categorical attribute, COOLCAT can be adopted to clustering.

(2) After the geometric center of each cluster is acquired, according to (3), the acceleration restriction is calculated.

(3) Initialize particle swarm: initialize the position for the searching point, and calculate the corresponding particle swarm.

(4) For every particle in the swarm, calculate its fitness value.

(5) For every particle, compare its fitness value and the one it has at the best position. If the latter is better than the former, renew P_{id}.

(6) For every particle, compare its fitness value and the group it has at the best position. If the latter is better than the former, renew $P_{gd.}$

(7) For every particle, according to position and speed's renewed (5) and (6), adjust its position and speed. The searching direction is non-decrease direction of acceleration.

(8) When the particle in the neighborhood can't meet the fitness requirement, and if there isn't clustered particle, choose un-clustered particle as the initial searching point, and jump to step3. When all the particles are marked, the algorithm ends.

4. The experiment validation

In the experiment the real data sets, Ecoli and Credit, in the UCI machine learning data sets, are used as the clustering objects. The detailed illustration of the data sets is as follows:

Table 1. Illustration of the data sets

UCI	Type	Element of attributes		Category	Instance	Missing data
		nominal	numeric			
ecoli	mixed	2	6	8	336	No
credit	mixed	9	6	15	690	Yes

Under the same environments the two data sets use K-Prototype and RPSOC respectively in the experiment. Therefore the accuracy of the clustering results is as follows in Fig.2.

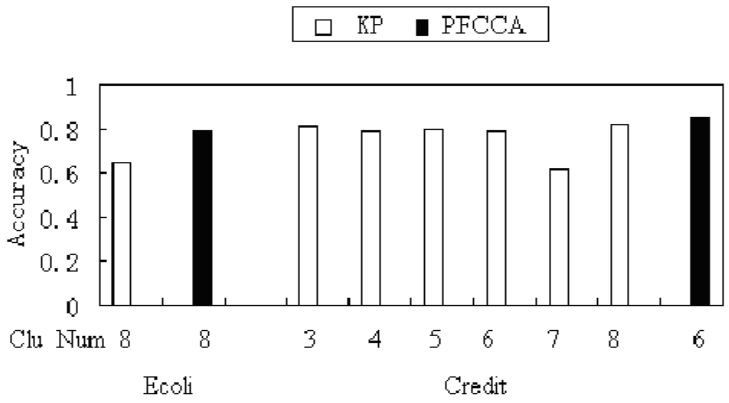

Fig.2. The accuracy of the clustering results

Accuracy of the algorithm can be defined as follows:

$$u = \sum_{i=1}^{k} \alpha_i \Big/ n \tag{7}$$

Here k is the number of the clustering, n is the number of the samples in the data sets and α_i is the number of the samples clustered accurately in ith cluster.

For Ecoli data set, contains two binary attributes, the clustering accuracy of RPSOC is higher than that of K-Prototype. For Credit, a data set with mixed attributes, clustering accuracies vary from clustering number to clustering number using k-prototype, because in the algorithm of K-Prototype it is necessary to preset the initial parameter. In such a case only by experiment can the best clustering result be attained. The algorithm of RPSOC, its average of clustering accuracies being 0.85, is better than k-prototype algorithm with the average being 0.77. Moreover, the RPSOC algorithm can automatically attain the number of clustering by means of guiding particles' to search for clustering through non-numeric attributes.

5. Conclusion

In this paper a completely new clustering algorithm processing data with mixed attributes is proposed, which aims to improve the speed and efficiency by means of restricting the searching direction of particle swarm in the gradient space. Compared with the traditional algorithm processing data with mixed attributes such as K-prototype and K-mode, it is an unsupervised self-learning process. Furthermore it doesn't need to preset parameter according to priori experience, which makes its results more directionally significant. The experiment in this paper effectively demonstrates the FPSOC algorithm can find clustering results of arbitrary shape and process data with noise and isolated point.

6. References

[1] Ganti V, Gehrke J, Ramakrishnan R. CACTUS: clustering categorical data using summaries[C]. In: SIGKDD Conf, 1999, 73-83.

[2] Guha S, Rastogi R, Shim K. Rock: a robust clustering algorithm for categorical attributes[J]. Information System, 2000, 25(5):345-366.

[3] Andritsos P, Tsaparas P, Miller R J. LIMBO: scalable clustering of categorical data[C]. In:9th Int Conf. On Extending DataBase Technology, 2004:531-532.

[4] Barbara D, Li Y, Couto J. Coolcat: an entropy-based algorithm for categorical clustering[C]. In: CIKM Conf, 2002,582-589.

[5] Darshit Parmar, Teresa Wu, Jennifer Blackhurst. MMR: An algorithm for clustering categorical data using rough set theory[C]. In Data & Knowledge Engineering, 2007,63(3):879-893.

[6] Agrawal R, Gehrke J, Gunopulos D. Automatic subspace clustering of high dimensional data for data mining applications[C]. In: SIGMOD Record ACM Special Interest Group on Management of Data, 1998,94-105.

[7] Procopiuc C M, Jones M, Aggarwal P K. A monte carlo algorithm for fast projective clustering [C]. In: SIGMOD Conf, 2002,418-427.

[8] Aggarwal C, Wolf J, Yu P. Fast algorithms for projected clustering[C]. In Proceedings of the1999ACM SIGMOD International Conference on Management of Data, ACM Press,1999,61-72.

[9] Aggarwal C, Yu P. Finding generalized projected clusters in high dimensional spaces[C]. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000,29:70-81.

[10] Huang Zhe-xue. Extensions to the k-means algorithm for clustering large data sets with categorical values [J].Data Mining and Knowledge Discovery, 1998,2(1): 283-304

[11] Huang Zhe-xue. Clustering large data sets with mixed numeric and categorical values [C]. Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore:World Scientific, 1997: 21-34

[12] Kennedy J, Eberhart R. Particle swarm optimization[C]. IEEE International Conference on Neural Networks. Piscataway: IEEE Service Center,1995,1942-1948.