Research and Discussion on the Novel Big Data Clustering Algorithm based on Probability Theory and Nash Game Theory

Haijun Liang[1]

[1] Hebei College of Industry and Technology,
Shijiazhuang,Hebei,China

Abstract. In this paper, we conduct research on the novel big data clustering algorithm based on probability theory and Nash game theory. Clustering algorithm is an effective method of data analysis, clustering algorithm is without any prior information of data clustering analysis of data and this kind of algorithm is also known as unsupervised learning methods. The Nash game theory and probability enhance the performance of the traditional clustering algorithm. The experiment result proves the feasibility of the combination. We set the schedule and prospect in the final part.

Keywords: Data Clustering; Probability Theory; Nash Game Theory; Experimental Analysis.

Introduction

Database from the very large scale data mined is used to extract information of interest. Clustering is an important tool of data mining. Clustering by establishing the mathematical model of the database is divided into different parts according to the data similarity, making data as similar as possible, within the class differences between objects. Different from general clustering algorithms for data mining clustering algorithm to deal with very large scale database, and types of data attribute is very much, so try to reduce the computational complexity of the algorithm. Clustering is a collection of physical or abstract objects grouping as the process of multiple classes of similar objects. Its purpose is to make the degree of similarity between individuals belonging to the same category as large as possible, and the degree of similarity between different categories of individuals as small as possible. In the field of machine learning, clustering is an example of no guidance of learning. Clustering algorithm is an effective method of data analysis, clustering algorithm is without any prior information of data clustering analysis of data and this kind of algorithm is also known as unsupervised learning methods. In many practical problems, due to the data itself without any prior analysis, the traditional clustering algorithms are sometimes not effectively clustering results. In the actual problem, sometimes we get a few data prior knowledge, including the class label and stronghold of classified constraints (such as constraint information in pairs). How to use these only a small amount of prior knowledge to no prior knowledge to a large number of data clustering analysis become a very important problem. A semi-supervised clustering is proposed for this kind of problem, it is using a small amount of a priori knowledge of the data to aid unsupervised clustering, so a semi-supervised clustering has gradually become the hot issue of clustering analysis [1-3].

The clustering results of algorithm, the purpose is to make as far as possible consistent with the known prior knowledge. Compared with a semi-supervised clustering algorithm, supervision and clustering algorithm more prior knowledge is required to ensure the effectiveness of the clustering results. This led to the objective function of the supervision and clustering algorithm and a semi-supervised clustering algorithm are quite different. Generally speaking, supervision and clustering objective function consists of two parts, including class not purity index and the number of clustering cluster [4]. Class not purity said clustering cluster contained in the non-mainstream tag data, the proportion of the so-called non-mainstream tag data refers to some data, it has the class label is not the most frequent in the clustering cluster of class labels, obviously, the class is not purity is lower, the better. In addition, the algorithms tend to clustering number smaller clustering results are obtained. Compared with a semi-supervised clustering, clustering need to provide supervision of prior information must be class labels, pair constraint information cannot be used to monitor the clustering algorithm. From the above analysis, a semi-supervised clustering analogy supervision cluster needs less prior information to many which is also much less.

Therefore, to deal with the mentioned drawbacks and disadvantages, we conduct theoretical research on the novel big data clustering algorithm based on probability theory and Nash game theory. In the figure one, we show the sample of the data classification and clustering result. The detailed analysis will be introduced in the following sections.

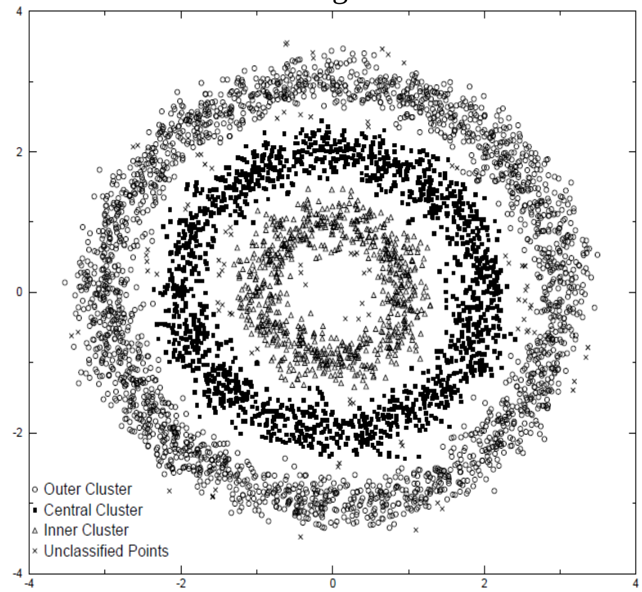

Fig. 1The Sample of Data Clustering and Classification

The Description of Our Proposed Algorithm

The General Analysis of Game Theory. The network is open and away from the equilibrium state. Between natural systems and external environment of material, energy and information exchange. In the case of a controlled environment, the network system of morphological evolution unceasingly, and the evolution process is not reversible. Network system is composed of lots of subsystem and the subsystem is composed of space time, distributed heterogeneous components. Network system of the overall macroscopic holds behavior by the nonlinear interaction between internal subsystems and components together and not by a single subsystem or component export. The system has the characteristics of self-organizing, local instability. In fact, the details of the network system is instability, even is very chaotic. But due to the nonlinear of the system and process irreversibility associated with local instability and unpredictability is dialectical unity which will eventually produces the stability of the system of macroscopic behavior. Energy to drive the stability of the research based on the ecosystem energy information, based on the particularity of the ecological system, gives a new definition method and the research methods and qualitative analysis was carried out on the system. The formula one shows the feature.

$$H\left(X\right)=-\sum_{i=1}^{N}p\left(x_i\right)\ln\left(p\left(x_i\right)\right)$$

(1)

A good system should not only have great energy flow, should also have higher information entropy and the entropy of the system is the sum of each entity information entropy. So, each entity has high redundancy information entropy is the necessary condition to improve overall redundancy. Entity can expand, shrink and output, or adjust to other entities in the output share, so as to achieve their own value, but these adjustments will be at the expense of the change of the whole system, in order to achieve the balance of each entity to pursue high entropy redundancy value, physical energy output can be seen as strategy, the utility function for each entity energy input function, make more game between each entity, because in the design of the energy of each entity restraining each other, each other policy changes will affect the effectiveness of each entity, so the game relation was established, through the game to seek the highest total entropy redundancy Nash equilibrium, provides the probability basis for the system of decision-making, guide policymakers orientation shown below in the following expressions.

$$f\left(x_i, x_j\right) = \max_{u \in X} f\left(x_i, x_j\right)$$

(2)

$$f\left(x_i, x_j\right) \geq \max_{u \in X} f\left(x_i, x_j\right)^2 - \min_{u \in X} f\left(x_i, x_j\right)^2$$

(3)

Concepts and Principles of Data Clustering. Spread neighbor clustering of another advantage is that it is the data form the similarity matrix of the symmetry, without any requirements for it expanded its range of application. However, for some of the data set itself has complex structure, spread neighbor clustering generally cannot receive reasonable clustering result. In this paper, the original neighbor propagation algorithm combined with a semi-supervised thought, heuristic to introduce known tag data or some constraints in pairs to adjust the similarity matrix. The similarity matrix is obtained by the new neighbor spread on the basis of the clustering, achieve the goal of improve the clustering performance. Adjustment based on the above two steps of preliminary results, based on the principle of the shortest path to the data contained in a priori information on the similarity of the global adjustment. If the data set is a data point and regulating connected respectively, and the data to the data points with the logarithmic stronghold of the similarity is greater than the sum of the initial similarity of logarithmic stronghold, is to adjust the data points of similarity for larger similarity. The algorithm of the basic idea is to cluster each object in the weighted average calculation, to assign each object in the database to the most similar clusters, repeated this operation, until the convergence criterion function even if the sum of square error to the degree of satisfaction. This is clearly aimed at numerical attribute data and attribute data symbols average power directly to objects in the cluster, and then to each object in the database to readjust.

$$E = \sum_{j=1}^{k} \left| p_i - AWM_j \right|^2$$

(4)

The Probability based Clustering. Internet platform service extension module design of the regulating function of the system, and the stable incentive analysis including energy adjustment, the adjustment of the system is not centralized management, the entity will adjust strategy according to their own and the stability of the system analysis of the complete evolution requirements, each entity will find their optimal stable strategy, the optimal state of the system is Nash equilibrium. Energy entity through the energy design mechanism, rely on to provide services to gain energy and at the same time various operations and use resources consumes energy, obviously, only provide a good service and minimum energy consumption can make energy accumulated entity, after a period of time, the system will maintain the high quality energy through the natural selection of energy entity, the energy of low energy entity will die from lack of energy. This energy entity will tend to high quality service, the evolution process of service. The process is shown below.

$$Q_{i,j} = \sum_{i=1}^{m} \sum_{j=1}^{n} F_{ij} / T$$

(5)

Grid environment, resource providers can be regarded as the producer, the application request as a consumer, they are two important factors of economic grid model. Computing economic model introduces the concept of economy to the grid resource management and it applies the principle of supply and demand in the market economy adjusted to owners and users of resources, to ensure the best interests of both sides to obtain. Grid economy model mainly includes: the grid resource request agent, grid middleware, domain resource management. Completely competitive market is the economic and social ideal, simple model, simply analyzes the supply and demand in the economy, while ignoring the individual supply and demand between mutual influence and restriction. Because of the complexity of the grid environment, and the size, and its direct competition and mutual restriction between resource providers is very powerful, so this article on the basis of general equilibrium theory, further research based on the theory of Nash equilibrium mechanism of resource management and scheduling which is shown below.

$$Q_{i,j} = \sum_{i=1}^{m}\sum_{j=1}^{n} F_{ij} / T + \sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij} / T - \sum_{i=1}^{m}\sum_{j=1}^{n} K_{ij} / T \qquad (6)$$

Algorithm makes the same kind of constraint data points were divided into the same class as much as possible, without similar constraints logarithmic stronghold in the end could not be divided into the same class. This change is the direct effects of the adjustment of the similarity matrix. In addition, from the formula, the representative value adjustment will also lead to optimum choice of value calculation iteration process change and formula suggests that represent the degree of value changes also affected by the optimum choice value. This loop iteration process shows that the similarity of the degree of change will eventually lead to all the data on behalf of the value of the selected value will be changed, and fitness has led to a great difference to the results of the final convergence algorithm.

Experiment and Result

In this part, we show the performance of our proposed methodology in the bid data environment. Algorithm fusion between closer to according to the traditional clustering algorithm type for many new algorithm to classify, layered algorithm based on grid and density algorithm thoughts or the idea of algorithm and so on may be reflected in an algorithm. The figure shows the result.

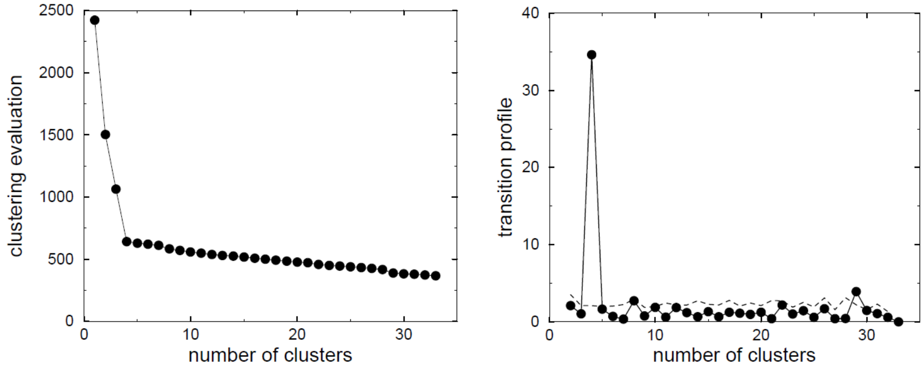

Fig. 2The Experimental Result for the Methodology

CONCLUSIONS

This paper conducts theoretical and numerical research on the novel big data clustering algorithm based on probability theory and Nash game theory. Clustering is a collection of physical or abstract objects grouping as the process of multiple classes of similar objects. Its purpose is to make the degree of similarity between individuals belonging to the same category as large as possible, and the degree of similarity between different categories of individuals as small as possible. The experimental result proves the effectiveness of the Nash game theory based methodology. In the future, we plan to conduct more related approached to polish the performance of the proposed methodology.

References

[1] Sancho-Asensio A, Navarro J, Arrieta-Salinas I, et al. Improving Data Partition Schemes in Smart Grids Via Clustering Data Streams[J]. Expert Systems with Applications, 2014, 41(13).

[2] Yuan F, Zhan Y, Wang Y. Data Density Correlation Degree Clustering Method for Data Aggregation in WSN[J]. IEEE Sensors Journal, 2014, 14(4):1089 - 1098.

[3] Lin, P., Huang, P., Kuo, C. H., & Lai, Y. H. (2014). A size-insensitive integrity-based fuzzy c-means method for data clustering. Pattern Recognition, 47, 5, 2042–2056.

[4] Prakash, J., & Singh, P. K. (2014). An effective hybrid method based on de, ga, and k-means for data clustering. Advances in Intelligent Systems and Computing, 1561-1572.