An Efficient Distributed Database Clustering Algorithm for Big Data Processing

Qiao SUN[1], Lan-mei FU[1], Bu-qiao Deng[1], Xu-bin Pei[2], Jia-song SUN[3,a]

[1] Beijing GuoDianTong Network Technology Co., Ltd, Beijing, China

[2] State Grid Zhejiang Electric Power Co., Ltd, Hangzhou, China

[3] E. E. Department, Tsinghua University, Beijing, China

[a]email: littlesmart@yeah.net

Keywords: Distributed big data processing, Distributed database, Data clustering, Depth neural network, K-means

Abstract. This paper proposes a distributed data clustering technique based on deep neural network. First, each record in the distributed database is taken as an input vector, and its characteristics are extracted and input to the input layer of the depth neural network. The weight of the connection is trained by BP algorithm, and the training of depth neural network output is realized by adjusting the weight. Finally, the data clustering results are judged according to the similarity of the current vector corresponding to the output data. Experimental results based on small-scale distributed systems show that this method has better test set accuracy than traditional k-means clustering method, and is more suitable for large-scale data clustering in the distributed environments.

1 Introduction

The distributed database system, which is composed of multiple computing devices connected by network, has become the mainstream database system. Distributed database has become an attractive method for a large number of users. In the distributed database, it is necessary to maintain the local control and decentralized management, but also to maintain the overall organization of the overall control and a higher level of collaborative management. This collaborative management requires the information between the various departments of enterprises can not only flexible exchange and sharing, but also unified management and use, enhance data reliability and availability. The data of a distributed database is stored on multiple sites that are geographically distributed. In these systems, data clustering technology is very important. In brief, the data clustering method based on distributed database improves the usability of data and reduces the cost of retrieval [1,2,3].

In this paper, a distributed data clustering algorithm based on depth neural network is proposed. Firstly, each record in the distributed database is taken as an input vector and its features are extracted and input to the input layer of the depth neural network. Finally, the data clustering results are judged according to the degree of similarity of the current vectors corresponding to the output layer to different data classes. The results show that the algorithm is effective and effective in the

training of the neural network. Compared with the traditional k-means clustering method, this method has better global optimization and robustness, so it is more suitable for large data clustering in distributed environment.

2 Big data analysis and mining functions

Big data analysis and mining functions are mainly reflected in the following areas:

Data can be associated with a class or concept, and a concept is often an overview of the overall situation of a data set that contains a large amount of data. Summary of data sets with large amounts of data is summarized and a concise, accurate description is described, which is called a class / concept description.

Association analysis is to find frequent patterns of knowledge from a given data set, also known as association rules.

Classification is to find a set of features (or functions) that can describe typical characteristics of a data set, so as to be able to classify and identify the attribution or classification of unknown data, that is, mapping unknown cases to one of the discrete classes. The classification pattern (or function) can be learned from a set of trained data (whose class is known) through the classification mining algorithm.

The difference between clustering analysis and classification prediction method is that the data used by the latter learning classification prediction model are known class attributes, belong to supervised learning method, and the data analyzed and analyzed by clustering analysis are none Category or without prior determination of category attribution.

The database may contain data objects that are not consistent with the general behavior or model of the data, which are referred to as isolated points. Most data mining methods discard outliers as noise or anomalies, but in some applications, such as automatic detection of various commercial frauds, small probability events tend to be more valuable than those that occur frequently. Isolated point data analysis is often called outlier mining.

Data evolution analysis is to describe the changing rules and trends of data objects that change with time. This modeling method includes concept description, contrast concept description, association analysis, classification analysis and time-related data analysis.

In order to achieve the above functions, many effective data mining techniques and algorithms have been proposed, such as C4.5, k-Means, SVM and Apriori [4,5,6].

3 Depth neural network principle

The neural network originates from a computational model proposed by McCulloch in 1943 to model the neural activity of living organisms [7]. In 1958 Frank Rosenblatt proposed a two-layer shallow neural network for pattern recognition of the perceptron model. However, the development of neural networks has been subject to significant resistance [8], because there are two key issues: Perceptor model cannot achieve XOR operation; then the computing power and can not meet the large-scale neural network computing needs [9]. In 1974, Werbos proposed the Back Propagation (BP) algorithm [10], which solves the problem that the neural network cannot perform XOR operation, and improves the training speed of the multi-layer network effectively. Even so, the depth of the neural network due to hardware limitations and theoretical complexity, in the ensuing period of time and did not get great progress. During this time, a variety of shallow neural networks have been developed, such as support vector machine (SVM) and other algorithms, the depth of the neural network by the cold. The depth of the neural network by the cold stop in 2006. Geoffrey Hinton put forward the concept of depth learning [11], which proved that depth neural network can characterize the features more deeply while enhancing the training efficiency of deep neural

network. Thus, setting off a deep research on the deep neural network.

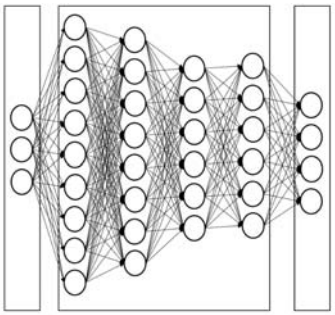

Fig.1 The structure of Depth Neural Network

As shown in Fig. 1, it is a deep neural network with three hidden layers. In each of the two adjacent layers, each node of each layer has a one-way connection with all the nodes of the other layer. Data input from the input layer, layer by layer to spread the next level. The activation function of the neural network is generally nonlinear, so that the depth neural network can fit any nonlinear function by changing the connection weight. That is, each node sums the output of all the nodes connected to the node according to its connection weight as the input of the node. The input is inputted to the next layer network through the activation function and finally the output of the output layer as the output of the entire network.

4 Depth Neural Network Clustering Algorithm Based on Distributed Database

The biggest advantage of depth learning is the ability to express a much broader set of functions in a more compact and concise way than a shallow neural network. That is, you can find simple functions that can express other functions through a composite of multiple layers of the network. When the processing object is a load curve, the partial-holonomic decomposition of the depth learning model can be applied. For example, in the first layer of the hidden layer can learn the overall profile of the load curve; and in the second layer can learn the load curve of the behavioral change point, in this layer can be based on the location of the mutation of the initial load Classification; at later levels, more features of the load curve can be learned, and these contours are further combined to detect more complex features [12]. In the depth neural network used in this paper, the parameters required for the whole network are the connection weight between every two layers. The activation function is the function that is determined at the beginning of network generation, so it does not occupy the number of parameters.

For the connection weights between the nodes of the deep neural network, the BP algorithm is used for training. BP algorithm for a given input and output training data, the first by the positive propagation from the input to be output, and then through the actual output and the correct output of the theoretical difference to be residual, and the output layer to the input layer according to the activation function and connection weight The residuals of each node and the ideal value are calculated. Finally, the weight of the connection between nodes is modified according to the residual of each node, and the training of the depth neural network output is realized by adjusting the weight. Closer to the theoretical output. Since the input layer input of the depth neural network extracts features from each input vector, the forward propagation through the network will be clustered according to the similarity of the current vector outputted by the output layer to different data classes.

5 Experiments and results

In order to test the performance of the above methods, we evaluated and tested in the existing network environment. Experimental configuration: 6 servers IBM-x3650M4-2U (2CPU, 6-core 12-thread Xeon E5-2620, 48G memory, 2T hard disk), the operating system is Ubuntu11.10 and Windows7, FTP services are pre-installed. Firstly, the k-means algorithm is compared with the 4-layer depth neural network clustering algorithm. The results are shown in Table 1, and it can be seen that the latter has obvious advantages over the former.

It can be seen from the experimental results that the increase of the number of layers does not cause the abrupt change of the parameter level of each layer because of the limitation of the number of parameters. From the present three structures, with the gradually increased number of layers. In other words, for such a deep neural network structure, with the limit of the total number of parameters, with the number of layers and the dimension of each hidden layer, we can get better clustering results. In order to maintain a better computational speed, the implicit layer dimension and the parameter magnitude are all limited. The results are shown in Table 2.

TABLE I. K-MEANS ALGORITHM AND 4-LAYER DNN DATA CLUSTERING

Algorithm	Network layer	Hidden layer dimension	Parameter magnitude (10^4)	Accuracy rate
K-means	0	0	10	44.8%
DNN	4	690	215	78.5%

TABLE II. DATA CLUSTERING RESULTS OF DIFFERENT NEURAL NETWORKS WITH DIFFERENT LAYERS AND HIDDEN LAYERS

Algorithm	Network layer	Hidden layer dimension	Parameter magnitude (10^4)	Accuracy rate
DNN	4	690	215	78.5%
DNN	6	532	214	81.7%
DNN	8	456	215	86.2%

6 Conclusions

In this paper, a distributed data clustering algorithm based on depth neural network is proposed. Firstly, each record in the distributed database is taken as an input vector and its features are extracted and input to the input layer of the depth neural network. Finally, the data clustering results are judged according to the degree of similarity of the current vectors corresponding to the output layer to different data classes. The results show that the algorithm is effective and effective in the training of the neural network. Compared with the traditional k-means clustering method, this method has better global optimization and robustness, so it is more suitable for large data clustering in distributed environment.

Acknowledgements

This research was financially supported by Science and Technology Project of the State Grid Corporation of China (SGZJ0000BGJS1500433) and the State Grid Information & Telecommunication Group CO., LTD. (SGITG-KJ-JSKF [2015]0003).

References

[1] Che Wujiang. Distributed database of data locking and consistency study [D]. Master's thesis, Hunan: Hunan University of Science and Technology, 2011
[2] SONG Chang-hong, LIU Yu-dong, ZHU Jie. Master-slave data consistency updating strategy based on message [J]. Computer Engineering, 2004, 30 (1): 92-94.
[3] H. Robert. A Majority Consensus Approach to Concurrency Control for Multiple Copy Databases [J]. ACM Trans. On Database System, 2003, 4 (2): 543-549.
[4] Wang Zhenwu, Xu Hui. Principle and implementation of data mining algorithms [M]. Tsinghua University Press, 2015.
[5] Liang Yasheng, Xu Xin, etc. Data mining principles, algorithms and applications [M]. Machinery Industry Press, 2015.
[6] Fan Donglai. Hadoop massive data processing: detailed technical solutions and actual combat [M]. People's Posts and Telecommunications Press, 2015.
[7] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity [J]. The bulletin of mathematical biophysics, 1943, 5 (4): 115-133.
[8] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain [J]. Psychological review, 1958, 65 (6): 386.
[9] Casper M, Mengel M, Fuhrmann C, et al. Perceptrons: An Introduction to Computational Geometry[J]. Anesthesiology, 1987, 75(3):3356-62.
[10] Werbos P. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science[J]. Ph.d.dissertation Harvard University, 1974, 29(18):65-78.
[11] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18 (7): 1527-1554
[12] Lin Jinbo.Clustering fusion and depth learning in the application of load pattern recognition [D]. Master's degree thesis, Guangzhou: South China University of Technology, 2014.