CHEMICAL ENGINEERING TRANSACTIONS
VOL. 46, 2015
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Peiyu Ren, Yancang Li, Huiping Song

ISBN 978-88-95608-37-2; **ISSN** 2283-9216
DOI: 10.3303/CET1546055

Research on Cloud Computing and Its Application in Big Data Processing of Railway Passenger Flow

Xiao Yong*[a], Cheng Ying[b], Fang Yanjun[a]

[a] Department of Automation, Wuhan University, Wuhan 430072, China,
[b] Information Centre, Wuhan University, Wuhan 430072, China.
xiao14@sina.com

Modern railway has a high speed, heavy load and intensive development trend. It not only brings the opportunities of railway transport capacity and volume, but also makes the data of various types of large-scale continuous growth. In the stage of the development of the railway by the production enterprise to the service enterprise, it is necessary to vigorously promote the application of cloud computing, and to plan the cloud computing framework. We study a parallel support vector machine model based on multi-level SVM, and realize the parallel algorithm in cloud computing environment. The algorithm divides the large training data set into a number of small training sets by Map, and then a new SVM is combined with these small training sets. Finally, the data is trained to be a new SVM by Reduce. At the end of the paper, we use the SVM parallel forecasting method to predict the passenger flow of China Railway, and compare the performance of the distributed with that of non-distributed algorithms. Experimental results show that the proposed algorithm has better effect than single machine algorithm in terms of time consumption and classification accuracy. With the increase of nodes, the time consumption is significantly shortened.

1. Introduction

Modern railway has a high speed, heavy load and intensive development trend. It not only brings the opportunities of railway transport capacity and volume, but also makes the data of various types of large-scale continuous growth. So it has brought a huge challenge to the information process of the railway. In the stage of the development of the railway by the production enterprise to the service enterprise, it is necessary to vigorously promote the application of cloud computing, and to plan the cloud computing framework (Lei Yun Wan (2011), Shi Xianliang (2011), and Zhou Xiang (2009)). Railway information can make train schedule with the change of passenger flow more targeted. The cloud platform can be used to build the booking system contains ticketing, refund, booking appointments and many other subsystems including technical support, maintenance and monitoring of massive ticket information (Wang Yili (2010) and Wang Juan (2009)). Therefore, the processing capacity of the railway passenger flow data in the cloud platform is particularly important.

Since MapReduce is an important method of big data mining algorithm, the domestic and foreign scholars have been abundant in the research of this aspect. MapReduce programming model based on cloud computing is a good solution to the problem of massive information processing in parallel, and the MapReduce model has the function of load balancing and automatic processing the node failure (Dean J C (2004) and Doulkeridis (2014)). Cheng Tao et al. (2006) and M.Y. Eltabakh (2011) developed the MapReduce parallel programming model to machine learning, and realized many kinds of machine learning algorithms. Kun Deng (2011) and C. Engle et al.'s (2012) report gave a algorithm of parallel SMO support vector machine, which is based on the sample set uniform and randomly divided into a number of data sets that are not too small. It is more effective for linear vector machine. Nasullah et al. (2011) and F. Färber, et al. (2012) used this idea to linear support vector machine. The experiment proved the effectiveness of the proposed idea for linear support vector machines. MapReduce parallel computing framework has been widely concerned by researchers because of its excellent performance and the simplicity of programming model in dealing with

Please cite this article as: Xiao Y., Cheng Y., Fang Y.J., 2015, Research on cloud computing and its application in big data processing of railway passenger flow, Chemical Engineering Transactions, 46, 325-330 DOI:10.3303/CET1546055

large-scale data. So, a lot of data processing algorithms based on MapReduce parallel framework have been implemented and achieved good performance(Yin Chuanye (2014) and Tang Duoyu (2013)).
In this paper, the application of cloud computing technology in railway large-scale data processing includes two aspects: on the one hand, based on the characteristics of the railway resource distribution, the paper proposes the method of using Hadoop cloud computing technology to analyze large data of railway passenger flow. This method can effectively improve the speed and efficiency of data analysis by constructing a cloud computing platform based on Hadoop. On the other hand, according to the data characteristics of the railway, we study a parallel support vector machine model based on multi level SVM, and realize the parallel algorithm in cloud computing environment. The algorithm divides the large training data set into a number of small training sets by Map, and then a new SVM is combined with these small training sets. Finally, the data is trained to be a new SVM by Reduce. At the end of the paper, we use the SVM parallel forecasting method to predict the passenger flow of China Railway, and compare the performance of the distributed with that of non distributed algorithms.

2. Cloud computing platform of railway passenger flow data based on hadoop

With the characteristics of railway passenger transport, the big data of railway passenger flow in the design of the cloud platform using distributed and hierarchical structure. As shown in Figure 1, it can be divided into 3 layers: data layer, model layer, and application layer.

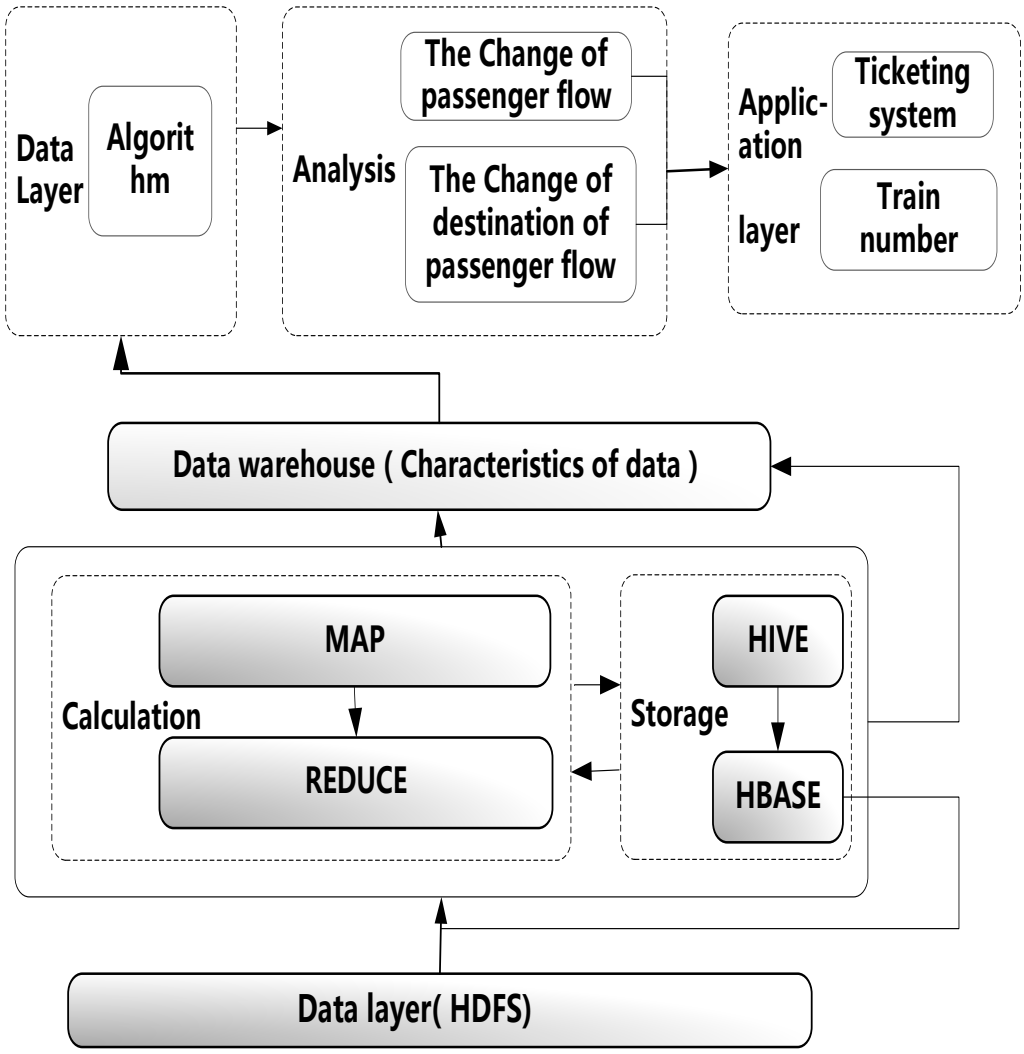

Figure 1: The architecture diagram of big data of railway passenger flow in cloud platform

(1) Data Layer
Railway passenger flow data including network data and business data. The data layer stores these data by HDFS Hadoop. And then, we can use Hbase, Hive and other data processing and management tools to dynamically generate Map-Reduce tasks. After that, through the high efficiency of the railway passenger flow data, processing results are still stored in file format in HDFS. At the same time, according to the specific business needs, we can export the required format.
(2) Model Layer.
The model layer uses the processing results of the big data of railway passenger flow which is processed in the data layer, and establishes the analysis model, such as the statistical model of the passenger's destination, the passenger's network booking model, the hierarchical demand model, the passenger flow forecast model, etc.. It can form an integrated analysis platform for the passenger's whereabouts and the network booking, so as to optimize the railway transportation system.
(3) Application Layer
The application layer analysis the results of model layer, such as the characteristics of the passenger's trajectory, and the concentration of the point of booking, so as to design a targeted sales program. It can effectively improve the efficiency of passenger transport of China's railway.

3. Parallel support vector forecasting model based on map-reduce

Support vector machine regression has two types including linear regression and non-linear regression. For linear regression, linear regression function is developed as follows:

$$f(x)=\omega x+b\,. \tag{1}$$

Assume training sample set D_n consist of n samples $(x_i,y_i)(i=1, 2,...,n)$, where $x_i \in X$, $y_i \in R$ are the best assessments. In order to guarantee flat of Eq. (1), we need to find a smallest w. Therefore, universal number of minimized Euclidean space is introduced and some technologies such as duality principle and lagrangian multiplier method are applied to find smallest w. Here, a regression function is obtained as follows:

$$f(x)=\sum_{i=1}^{n}\left(a_i-a_i^*\right)\left(x_i\cdot x\right)+b \tag{2}$$

The basic ideal of non-linear support vector machine regression is demonstrated in the following. Firstly, data can be mapped into high dimensional feature space by a non-linear mapping. Secondly, linear regression can be implemented in this space. Therefore, the linear regression in the high dimensional feature space can correspond to non-linear regression in the low dimensional input space, which can be achieved through core function $k\left(x_i,x_j\right)=\Phi\left(x_i\right)\Phi\left(x_j\right)$. Thus, the following equation can be acquired:

$$\begin{cases}\max -\dfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_i^*-a_i\right)\left(a_j^*-a_j\right)k\left(x_i,x_j\right)-\varepsilon\sum_{i=1}^{n}\left(a_i^*+a_i\right)+\sum_{i=1}^{n}y_i\left(a_i^*-a_i\right)\\ s.t.\sum_{i=1}^{n}a_i^*=\sum_{i=1}^{n}a_i,\\ a_i^*,a_i\in[0,\mathrm{C}],i=1,2,...,n\end{cases} \tag{3}$$

Where constant C >0 is named penalty factor. If C is big, fitness bias has bigger penalty. ε is biggest bias that regression can be allowed. Then, regression function can be demonstrated as follows:

$$f(x)=\sum_{i=1}^{n}\left(a_i-a_i^*\right)k\left(x_i,x\right)+b=\sum_{i=1}^{n}\alpha_i k\left(x_i,x\right)+b \tag{4}$$

Parameter b can be obtained from the following equation.

$$b=\begin{cases}y_l-\sum_{i=1}^{n}\left(a_i^*-a_i\right)k\left(x_i,x_l\right)+\varepsilon, if a_l\in[0,\mathrm{C}]\\ y_l-\sum_{i=1}^{n}\left(a_i^*-a_i\right)k\left(x_i,x_l\right)-\varepsilon, if a_i^*\in[0,\mathrm{C}]\end{cases} \tag{5}$$

$$K(x,y)=\exp[-(x-y)^2/\sigma^2] \tag{6}$$

Different from the traditional single model method, the parallel multi model method can be aimed at different clustering to select different kernel parameters for each sub model. Because kernel of Gauss is a typical local function. So, the distance between the data and the test point is getting closer, the influence on function value is higher. We can determine the kernel width parameter according to the average distance between the data and the clustering center. It is used to ensure that the data in the cluster has a greater impact on the value of the kernel function, while the data in the non cluster has little effect on the value of the kernel function. The calculation formula is as follows:

$$\sigma_i=\sqrt{\sum_{j=1}^{l}\left\|x_{i,j}-c_i\right\|_2/l} \quad i=1,\cdots,k;\ j=1,\cdots,l \tag{7}$$

Among them: c_i is the center of i th cluster, $x_{i,j}$ is the of j th sample data of i th cluster, l is the sample number in the i th cluster.

In this paper, the large scale training data set of SVM algorithm is divided into several small training sets by Map, and then a new SVM model is combined with these small training sets. That is to say, the α ($0<\alpha_i<C$) corresponds to the sample (x_i,y) as the input of Reduce. In the cloud computing environment, the training data set is automatically completed by the Map-Reduce system. It is based on the user's Map input block size, the big data set is divided into a number of small training sets that are not greater than the block size of the Map input. Then, small training sets are the input for Map task, and the Map tasks are independent and parallel execution. After each Map task completed the training of their small training set, the sample data of the support vector machine is used as the output of Map.

4. Simulation experiment and result analysis

Due to the length of the article and the author's limited energy, this paper only carries on the deployment of the Hadoop cluster of the railway passenger flow data, and experiments are applied to demonstrate the effectiveness and feasibility of the proposed method and the design of the platform. In order to verify the performance of the high performance computing system under the background of big data, this paper uses the Hadoop cluster environment is made up of 8 SYSTEM X3850 IBM server, each server is a quad core PC, each core as a Hadoop node. One of them is NameNode, the other is the DataNode. In addition, in order to compare the performance of the benchmark, we are running a contrast experiment in non distributed mode. After that, we use the fast Fourier-Transform algorithm to generate the test data sets of 3 different sizes: small data sets (S, 100MB), medium data sets (M, 300 MB) and large data sets (L, 500MB). We take different data sets as the sample data set of the parallel SVM algorithm, and compare with the non distributed SVM algorithm. Experimental results are shown in figure 2 and figure 3.

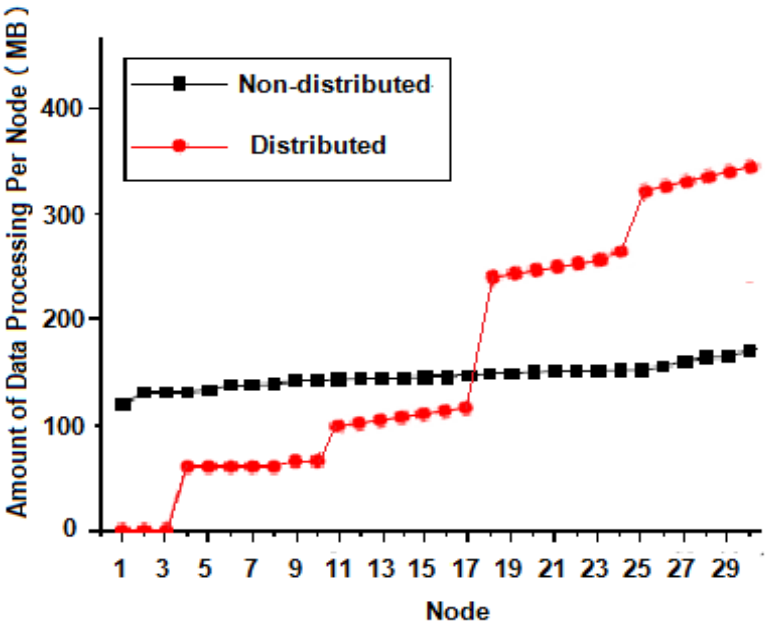

Figure 2: Data distribution with different base columns.

As can be seen from Figure 2, the amount of data processed by the two SVM algorithms in a same node dimension is different. The amount of data processed by parallel (distributed) SVM algorithm is less than traditional (non distributed) SVM algorithm. However, with the gradual increase of nodes, the amount of data processed by the two algorithms have gradually produced a difference. Although with the increase of nodes, the two algorithms can deal with more data. But the parallel (distributed) SVM algorithm is more obvious and more outstanding.

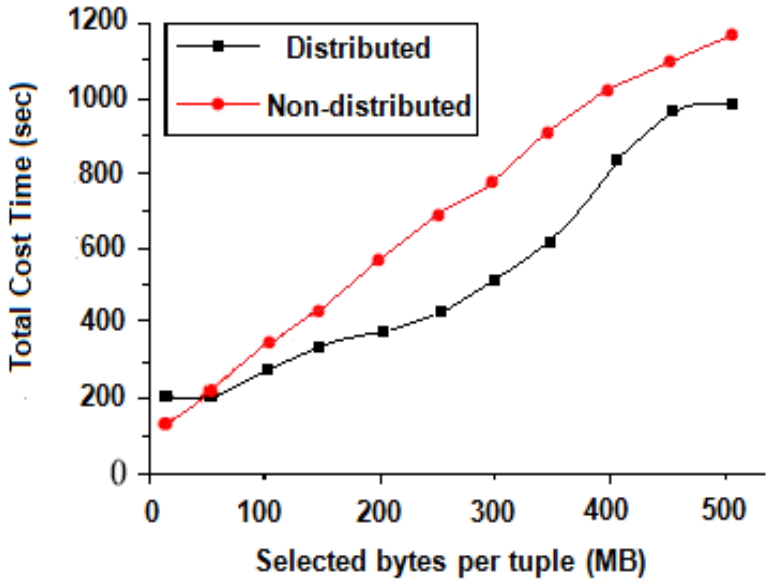

Figure 3: Performance evaluation of Map-Reduce execution with different base columns

Can be seen from figure 3, the time consuming of the computation of the two kinds of SVM algorithms in the same data dimension is also different. When dealing with small sample data sets, the difference between the two algorithms is not significant. And with the increase of the number of sample data, the time consuming of operation has gradually produced the difference. Obviously the computing speed of the parallel (distributed) SVM algorithm is faster.

5. Conclusions

In this paper, a distributed cloud computing method is proposed to improve the speed and efficiency of big data analysis of railway passenger flow data. We analyze the problem of low efficiency of the traditional data analysis method in the big data of railway passenger flow. Then, we propose a distributed cloud computing method based on Hadoop, design a large data cloud computing platform based on Hadoop, and realize the big data processing function of the parallel SVM passenger flow in the cloud platform. Experiments show that the method proposed in this paper is effective and feasible. In order to further study the application of Hadoop in railway passenger flow data analysis, this paper gives the direction of application.

References

Alham N.K., Li M.Z., Hammoud S., Liu Y., Ponraj M., 2010, A Distributed SVM for Image Annotation [C]. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010: 2983-2987.

Chu C.T., Kim S.K., Lin Y.A., 2006, Map-Reduce for Machine Learning on Multicore [C]. NIPS'06, 281-288.

Dean J., Ghemmawat S., 2004, MapReduce: Simplied data processing on large clusters [C]. OSDI* 04: Proceedings of the 6th Sympesium on Operating System Design and Implementation. New York: ACM Press, 137-150.

Doulkeridis C., Nørvåg K., 2014, A survey of large-scale analytical query processing in map-reduce [J]. VLDB J. 23 (3) 355–380.

Eltabakh M.Y., Tian Y., Özcan F., Gemulla R., Krettek A., McPherson J., 2011, Cohadoop: flexible data placement and its exploitation in hadoop [J], Proc. VLDB Endow, 4 (9): 575–585.

Engle C., Lupher A., Xin R., Zaharia M., Franklin M.J., Shenker S., Stoica I., 2012, Shark: fast data analysis using coarse-grained distributed memory [J]. in: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, 689–692.

Färber F., Cha S.K., Primsch J., Bornhövd C., Sigg S., Lehner W., 2012 , Sap hana database: data management for modern business applications [J]. SIGMOD Rec. 40 (4): 45–51.

Kun D., Yih L., Perera A., 2011, Parallel SMO for Training Support Vector Machines [EB/OL] [2011-01-17]. Http: / /www.geocities.com/asankha / Contents / Education /APC-SVM.pdf.

Lei Y.W., 2011, Cloud computing technology, platform and application [M]. Beijing: Tsinghua University press.

Shi X.L., 2011, Supply chain management [M]. Beijing: Machinery Industry Press.

Tang D.Y., Cao X.H., 2013, Parallel algorithm designing and application of weighted Voronoi diagram using MapReduce programming mode [J]. Application Research of computers, 30 (5): 1410-1412.

Wang J., Hong C.Y., Shen Z., 2009, Design of EC Website Search Engine Based on Compass [J]. Modern computer based on, (2): 129-131.

Wang Y.L., Yang X.J., 2010, Full-text Retrieval System Research and Design Base on Compass [J]. Coal technology, (6): 157-159.

Yin C.Y., Zhang Y., Wu C.Z., 2014, Research on Page-Rank algorithm optimization based on MapReduce [J]. Computer application research, 31 (2): 431-434.

Zhou X., Wang L.F., Jiang Z.J., Zhang Y., 2009, Design of enterprise information portal search engine based on Lucene [J]. Microprocessor, (8): 62-63.