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, Yanchang Li, Huiping Song

ISBN 978-88-95608-37-2; **ISSN** 2283-9216
DOI: 10.3303/CET1546062

Application of Parallel Particle Swarm Optimize Support Vector Machine Model Based on Hadoop Framework in the Analysis of Railway Passenger Flow Data in China

Wang Xun*[a], Yingbo An[b], Rong Jie[b]

[a] Human Resources Department, Hebei Software Institute,China
[b] Information Management and Engineering Department, Heibei Finance University, China
859071566@qq.com

In recent years, the development of high-speed railway industry in China is very rapid. However, the development of Chinese high speed railway cannot be further improved without basic research. The passenger flow is the basis and foundation to build high-speed railway. Therefore, to establish a set of analysis method for big data forecasting of railway passenger flow has great theoretical value and practical significance. In this paper, a parallel particle swarm optimization algorithm which is based on Hadoop framework is proposed, which can effectively avoid the particle swarm algorithm falling into local extreme value. Parallel particle swarm optimization algorithm is used to optimize the parameters (C, σ^2) of SVM. Taking into account the each solution of particle swarm adaptation value is to go through the quadratic optimization process of support vector machine, we use the particle swarm optimization in parallel computing to complete the rapid prediction of big data. Experimental results show that the algorithm has good performance and high accuracy, which proves the validity of the algorithm.

1. Introduction

In recent years, the development of high-speed railway industry in China is very rapid. China's high-speed railway has the characteristics of the most comprehensive system technology, the most integrated capacity, the longest running mileage, and the highest operating speed. However, the development of Chinese high speed railway cannot do without basic research. The passenger flow is the basis and foundation to build high-speed railway, reasonable organization of passenger flow is the key to play the benefits of high speed railway. Therefore, to establish a set of analysis method for big data forecasting of railway passenger flow has great theoretical and practical significance.

According to the limitation of different time, the forecast of passenger flow can be divided into long-term and short-term forecast. At present, the research on the forecast of passenger flow is mainly concentrated in the annual and monthly forecast, while the short-term forecast of passenger flow is less, especially the daily passenger traffic forecast based on real time data is very rare. To sum up, these models and methods can be divided into three categories that include the model based on time series, modern mathematical algorithm, and characteristics of passenger flow. The forecasting method based on time series includes regression analysis, exponential smoothing, time analysis and grey model (Zheng Yan (2010), Wang Ting (2007),) Dong Bing (2010), Wang Jinbiao (2003), and Wang Fang (2007)), etc.. On the other hand, the forecasting model based on the modern mathematical algorithm includes many methods such as neural network, support vector machine and wavelet theory. Each of these methods has its advantages and disadvantages, and it is based on the analysis of a large number of historical data (Huang Darong, Song Jun (2006)). There are many applications of the forecasting model which is based on modern mathematical algorithm. A passenger flow forecasting model was based on support vector machine that was proposed by Huo Baoshi, Zhang Xiuyuan (2001). They used the actual data to verify the model in precision, convergence time, generalization ability. Zhang Li, Xie Zhongyu and Chen Kai (2011) used the method of small data to identify the time series of passenger flow. They built an ALOMM forecasting model, which achieved good results.

Please cite this article as: Xun W., An Y.B., Jie R., 2015, Application of parallel particle swarm optimize support vector machine model based on hadoop framework in the analysis of railway passenger flow data in china, Chemical Engineering Transactions, 46, 367-372
DOI:10.3303/CET1546062

With the gradual realization of the modernization of China's railway, the amount of data generated by the railway passenger flow is increasing. So we need to use a fast and efficient intelligent computing method to deal with big data. Particle swarm optimization algorithm has been widely used in function optimization, neural network training, pattern classification, fuzzy system control and other fields (Zhang C (2000), Mendes F (2002) and Juang C (2004)). In the pursuit of optimal particle process, particle swarm optimization gets closer to the optimal particle, and its speed is getting smaller. So the particle swarm has a strong convergence, and it is easy to fall into local minimum point. Some improved algorithms of PSO were proposed. For example, SHI Y (1998 and 2001) et al. proposed the improvement of inertia weight and the Fuzzy adaptive PSO algorithm, the PSO algorithm for adaptive mutation proposed by Lv Zhensu et al. (2004), and an adaptive particle swarm optimization algorithm proposed by Yasuda et al (2003).

In this paper, a parallel particle swarm optimization algorithm which is based on Hadoop framework is proposed, which can effectively avoid the particle swarm algorithm falling into local extreme value. Parallel particle swarm optimization algorithm is used to optimize the parameters (C, σ^2) of SVM. Taking into account the each solution of particle swarm adaptation value is to go through the quadratic optimization process of support vector machine, we the use the particle swarm optimization in parallel computing to complete the rapid prediction of big data. The tracking process of the particle swarm is completed in various relative independent parallel processes, which can ensure the diversity and the speed of the individual populations. In this paper, we use the support vector machine algorithm which is based on parallel particle swarm optimization to establish the forecasting model of railway passenger flow. Experimental results show that the algorithm has good performance and high accuracy, which proves the validity of the algorithm.

2. Cloud Computing

Cloud computing is based on the increased development of in Internet related services, and the use and delivery models. It is the result of the development of parallel computing, distributed computing and grid computing. The distribution of computing tasks constitutes a large number of computer resource pool. According to the specific requirements, we can make a variety of application systems to obtain computing power, storage space and information services. At the same time, cloud computing has the characteristics of data security, reliable, scalable, strong, large scale, low price and so on. According to different services, cloud computing is divided into SaaS (software as service), PaaS (platform as a service) and IaaS (infrastructure as a service). Then, the data is the center of cloud computing. It has a unique technology in the parallel data processing, programming model and virtualization.

3. SVM Model

Linear separable SVM is to obtain the best generalization performance by maximizing the classification margin, and in common condition, it usually can't get the ideal state. Take the positive relaxation factor into the optimization problem, we can describe the optimization problem as follows:

$$\min_{\omega}\left(\frac{\|\omega\|^2}{2}+C\sum_{i=1}^{n}\xi_i\right) \tag{1}$$

S. t $y_i[(\omega\cdot x_i)+b]\geq 1-\xi_i;\ i=1,...,n;\ \xi_i\geq 0, i=1,...,n$

Due to various parameters as ω, b and ξ_i, it is difficult to get the solution of the optimization problem. Convert the optimal hyper plane according to the Lagrange method, and we can get the following description:

$$\min_{\lambda}=\left(\frac{1}{2}\sum_{i}\sum_{j}\lambda_i\lambda_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{n}\lambda_i\right) \tag{2}$$

S. t $\sum_{i=1}^{n}\lambda_i y_i=0, \qquad 0\leq\lambda_i\leq C$

Where λ_1 is the Lagrange multiplier. According to KKT conditions.

The input sample data set of the SVM is n, then the training sample will be:

$$s_n=((x_1,y_1),(x_2,y_2),...,(x_n,y_n)),\ yi\in\{-1,1\} \tag{3}$$

For linearly separable training samples, SVM could find the hyper plane with the maximum Euclidean distance and the nearest training sample. And for the non-separable training samples, the total error rate can be expressed with slack variables N_i. Calculation of the hyper plane is equal to the solution of the basic optimization problems as follows:

$$\min V(\omega, b, \xi) = \frac{1}{2}\omega^T \omega + c\sum_{i=1}^{n} \xi_i \tag{4}$$

S. t $\forall_{i=1}^{n} : y_i[\omega^T x_i + b] \geq 1 - \xi_i$, $\forall_{i=1}^{n} : \xi_i > 0$

Give the typical kernel function as follows:

$$K(x_i, x_j) = (x_i^T x_i + 1)^d \text{ and } K(x_i, x_j) = \exp(-r\left\|x_i - x_j\right\|^2) \tag{5}$$

Then, the decision function is obtained as

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} a_i y_i K(x, x_i) + b\right) \tag{6}$$

From the computation process of SVM, the different values of ε in non-sensitive loss function, penalty coefficient C and σ^2 in radial basis function will lead to the different support vector regression model. Therefore, in this paper, we select approximate optimization of parameter sets (C, σ^2) based on PSO by controlling the value of ε to construct task scheduling algorithm of PSO -SVM.

4. Parallel PSO Model

The core idea of parameter optimization in particle swarm optimization is treating the two-dimensional vector as the position of the particle and set a reasonable objective function at the same time. When each particle is searching by location, the purpose is to minimize or maximize the objective function and determine its historical best point in group or domain.
The objective function is set to the mean square error function:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{7}$$

Among them, y_i is the true value, $\hat{y}_i$ is the predicted value.

Then writing C as $x = (x_1, x_2)$. It consists of particles in groups. Then the position of particle i can be expressed as $x_i = (x_{i1}, x_{i2})$, while velocity of particle i is $v_i = (v_{i1}, v_{i2})$, its historical best point can be written as $p_i = (p_{i1}, p_{i2})$ and the whole best point can be written as $p_g = (p_{g1}, p_{g2})$. Then the position and velocity of the particle will change with the following equation:

$$v_{ij}^{(t+1)} = wv_{ij}^{(t)} + c_1\delta_1(p_{ij}^{(t)} - x_{ij}^{(t)}) + c_2\delta_2(p_{gj}^{(t)} - x_{ij}^{(t)})$$

$$x_{ij}^{(t+1)} = x_{ij}^{(t)} + v_{ij}^{(t+1)}, \quad j = 1, 2$$

Among them, c_1 and c_2 are known as learning factors and always equal to 2. δ_1 and δ_2 are pseudo random number whose interval is $[0,1]$. w is the inertia weight, its value will influence the exploration ability and explore ability of the algorithm. We make the value of the time-varying as weights and hypothesis $w \in [w_{\min}, w_{\max}]$, $Iter\_\max$ is maximum number of iterations.

$$w_i = w_{\max} - \frac{w_{\max} - w_{\min}}{Iter\_\max} * i$$

,

Among them $[w_{\min}, w_{\max}] = [0.1, 0.9]$.

Now we use the idea of cross validation to optimize the PSO-SVR model in order to find a more reasonable set of parameter (C, σ^2), so the model's error is smaller.

The MapReduce distributed programming model is firstly proposed by Google laboratory. The method is mainly used for parallel computing of large-scale data sets. It is a kind of programming model based on function. The operation of massive data set is abstracted into two sets of Map and Reduce, and the bottom layer is encapsulated, which greatly simplifies the implementation of the program. In MapReduce computing mode, the user is required to provide the Map function and the Reduce function to achieve the mapping and the reducing process. This is calculated by the Map and Reduce function for a set of input keys, which calculated another set of keys:

$$\begin{cases} Map: (x_{i1}, v_{i1}) \to list(x_{i2}, v_{i2}) \\ Reduce: (x_{i2}, list(v_{i2})) \to list(x_{i3}, v_3) \end{cases} \tag{8}$$

The Map function receives a set of input keys (x_{i1}, v_{i1}) that is processed to produce a set of intermediate values (x_{i2}, v_{i2}). Then the MapReduce function library sends the output values to the Reduce function. The output value $list(v_{i2})$ is the value of v_{i2} which is corresponding to key value of all the same x_{i2}. By merging the set of middle key value, it can form a new set of key value which is $list(x_{i3}, v_3)$. Repeat this process, we can get the set of final key value which is $list(x_{in}, v_n)$.

5. Experiment and Result Analysis

Construction of Hadoop cluster platform. We build and configure the Hadoop platform on the 8 PC cluster in the laboratory LAN. The Hadoop version is Hadoop 0.20.2, the operating system is Ubuntu 10.4, and the hardware environment of the PC is the same as Dual-core (R) CPU E6300 @2.8 GHz Apache ADAT Pentium, 2 Gbit Hitachi memory, and 320 Gbit hard drive.

Experiment 1:

First of all, we divide the railway passenger flow data of our country in 1987-2010 into several blocks that is according to the time dimension, so as to form different new data sets. We take each PC as a Hadoop node. So , there are 8 nodes.We take the historical data of the first nine years as a set of training, and forecast the passenger flow of the last five years. The performance of the parallel PSO-SVM algorithm and the traditional PSO-SVM algorithm at different nodes, as shown in the following figure:

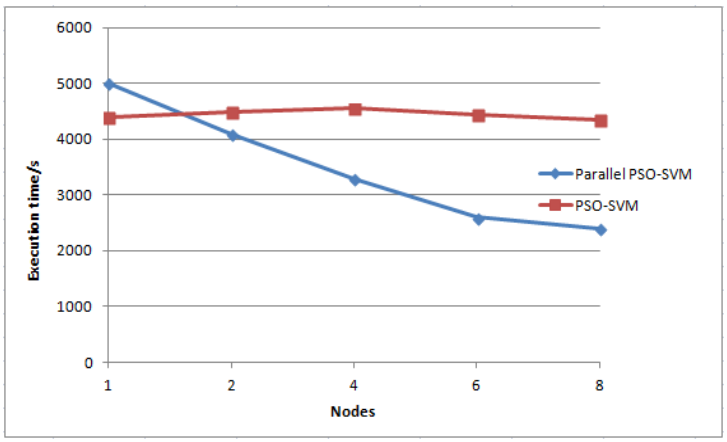

Figure 2: The Execution Time Comparison Figure

It can be seen in Figure 2 that the operation time of the sample processed by the two kinds of SVM algorithms is different. When we use only one machine to perform, due to the existence of communication time consuming, based on the MapReduce of the parallel SVM algorithm is more time-consuming. But when the machine group is expanded, it can be seen that the execution time decreases rapidly, which is much less than that of the single machine. When the nodes are 6 or 8, the execution time is basically stable, which is related to the task size. When the number of the machine has been satisfied with the MapReduce distribution, the machine will not have a significant effect on the execution time.

Experiment 2:

We use the data of our country's railway passenger traffic in 1987-2005 as the training data set of PSO-SVM, and use the 2006-2010 data as a contrast data set. We use the parallel PSO-SVM algorithm, the traditional PSO-SVM algorithm, SVM algorithm and GM (1,1) algorithm for forecasting, the results are shown in the following figure:

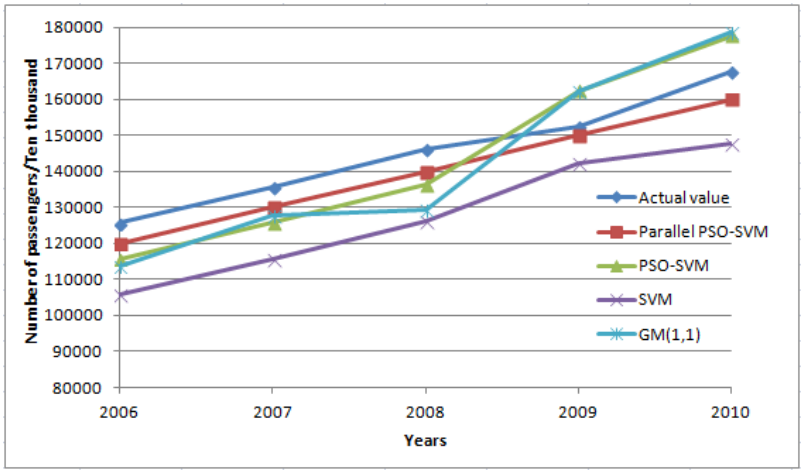

Figure 3: Comparison of forecasting results

From Figure 3, we can see that although the process of optimizing the SVM parameter increases the computation time of the algorithm, but the parallel processing method of the particle swarm optimization SVM algorithm can effectively improve the accuracy of prediction. The proposed method makes the optimization operation of the loop executes in parallel, so the increased computation time is very short. In this paper, compared with other forecasting methods, the accuracy of prediction is improved obviously.

6. 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, and realize the function of the parallel PSO-SVM traffic prediction in the cloud platform. Experiments show that the method proposed in this paper is effective and feasible.

Acknowledgements

This research is supported by the Youth Fund Project of Hebei Education Department (QN2015161) and Youth Fund Project of Hebei Education Department (QN2015151).

References

Dong B., 2010, Civil aviation passenger flow data prediction based on multivariable grey model [J]. Journal of Civil Aviation Flight University of China, (1): 21-23.

Huang D.R., Song J., 2006, Forecasting Model of Traffic Flow Based on ARMA and Wavelet Transform [J]. Computer Engineering and Applications, (36): 191-194, 224.

Huo B.S., Zhang X.Y., 2001, Application of Fuzzy clustering analysis principle in passenger flow forecast [J]. The Journal of Quantitative and Technical Economics, (12): 90-93.

Juang C., 2004, A hybrid of genetic algorithm and particle swarm optimization for recurrent network design [J]. IEEE Trans on Systems, Man and Cybernetics-Part B: Cybernetics, 34(2): 997-1006.

Lv Z.S., Hou Z.R., 2004, Particle Swarm Optimization Algorithm with Adaptive Mutation [J]. ACTA ELECTRONICA SINICA, 32(3): 416-420.
Mendes R., Cortez P., Rocha M, Neves J., 2002, Particle swarms for feed-forward neural network training [C]// Proc of Int Joint Conf on Neural Networks. Honolulu: IEEE Computer Society, 2002: 1895-1899.
Shi Y., Eberhart R., 1998, A modified Particle swarm optimizer [C]// Proc IEEE Int Conf on Evolutionary Computation. Anchorage: IEEE Press, 1998: 69-73.
Shi Y., Eberhart R., 2001, Fuzzy adaptive particle swarm optimization [C]// Proc of IEEE Conf on Evolutionary Computation. Piscataway: IEEE Service Center, 2001: 101-106.
Wang F., 2007, Study on short term forecasting method of railway passenger transportation [D]. Beijing Jiaotong University.
Wang J.B., Feng Z.C., 2003, Revenue management forecasting model based on the excited level of passenger flow [J]. Journal of Civil Aviation University of China, 21(S2): 63-66.
Wang T., 2007, Modeling and Prediction of Civil Aeronautic Passenger Capacity By Using ARIMA Models [J]. Journal of WUYI University (Natural Science Edition), 21(1): 38-42.
Yasuda K., 2003, Adaptive particle swarm optimization [C]// Proc of IEEEE Int Conf on System, Man, Cybernetics. Indianapolis: IEEE Press, 2003: 1554-1559.
Zhang C., Shao H., Li Y., 2000, Particle swarm optimization for evolving artificial network [C]// Proc of IEEEE Int Conf on System, Man, Cybernetics. Piscataway: IEEE Service Center, 2000: 2487-2490.
Zhang L., Xie Z.Y., Chen K., 2011, Local region short-term traffic flow forecasting model based on chaotic theory [J]. Journal of Heilongjiang Institute of Technology, 2011, 25(2): 52-54.
Zheng Y., 2010, ARIMA adjustment and regression analysis on time series in aviation industry [J]. Journal of Qiqihar University (Natural Science Edition), 26(03): 82-84.