International Journal of Grid Distribution Computing
Vol.8, No.1 (2015), pp.33-40
http://dx.doi.org/10.14257/ijgdc.2015.8.1.04

Improved SVM in Cloud Computing Information Mining

Lvshuhong

(ZhengDe polytechnic college JiangSu NanJing 211106)
1713754023@qq.com

Abstract

How to have a better mining and use of information in the cloud computing environment constitutes the direction of current research in the field of cloud computing; this paper introduces support vector machine (SVM) concept in the cloud computing data mining, introduces a penalty factor in the SVM, and improves SVM data mining algorithms. The constructed Map / Reduce model by the concept of featured multi-tree conducted the validation of the model. Simulation results show that data mining methods of this model have effectively improved the accuracy and the time of information mining, hence have some practical significance.

Keywords: *data mining; penalty factor; Map / Reduce model*

1. Introduction

The growing exponentially cloud computing data no doubt poses a severe challenge to cloud computing data mining [1]. In the cloud computing, the data mining is based on a comprehensive analysis of different cloud nodes; people find the value of the information in the process of information mining, hence conduct processing information in cloud computing environment and provide a favorable data analysis for decision-making departments. How to dig out the information needed for the reader out of vast amounts of data in cloud computing environment, is a realistic problem indeed to be resolved.

At present, the commonly used data mining algorithms are neural networks, decision trees and Bayesian networks [2]; most methods conduct data mining under distributed environment, hence these data mining algorithms take up large storage space, increase network load,and extend the response time [3-4]. Literature [5] proposed a data mining algorithm based on dynamic cloud computing model, and simulation results show that the algorithm has certain advantages in terms of handling massive amounts of data. Literature [6] proposed an efficient data mining framework on Hadoop, using the database to simulate the chain structure, managing the excavated knowledge. Literature [7] proposed a cloud computing network health assessment method based on support vector machine. Simulation results show the feasibility of the method. Literature [8] by using the support vector machines in coal industry under the cloud computing conditions, effectively conducted mining and analysis for coal data and achieved some results.

2. Related Knowledge

2.1 Support Vector Machines

Support vector machine (referred to as SVM) is a machine learning method based on the VC theory, whose basic idea is to rely on a hyperplane as a decision, and make the maximum distance between the positive and negative modes. Relying on the optimal classification surface, SVM has been built up; the optimal classification surface has correctly separated the classified surfaces, while maintaining the maximum spacing between classification surfaces; the vector closest to the optimal classified hyperplane becomes

ISSN: 2005-4262 IJGDC

support vector.

The optimal partition hyperplane is calculate and described as a conditional extremum problem, obtaining the optimal solution by Lagrange function saddle point:

$$Lp = \frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i - \sum_i a_i\{y_i(\Phi(x_i)\cdot w + b) - 1 + \varepsilon_i\} - \sum_i \mu_i \varepsilon_i \quad (1)$$

By kernel function theorem, the optimal solution must satisfy the following conditions

$$w = \sum_i a_i y_i \Phi(x_i) \quad (2)$$

$$0 \le a_i \le C, \sum_i a_i y_i = 0 \quad (3)$$

In equation (2), under normal circumstances, the default value of a_i is 1, the result of w is nonzero of x_i. The support vector machine takes the training data set as a unit to describe the features of unit of data; in aspects of making the equivalent classification and segmentation data sets, vector machine occupies a very small part.

2.2 Data Mining in the Cloud Computing Environment

Currently, data mining in the cloud computing environments relies on web logs as a study, by Web data mining techniques to understand and analyze the daily behavior of the user; log mining in the cloud computing environment is divided into: (1) classification analysis, mainly conducts classification and analysis of user behavior characteristics under cloud computing conditions; (2) association rules analysis, to analyze the user pages behavior under cloud computing environment; (3) frequent sequence access analysis: to analyze users frequently access under cloud computing condition, hence obtain access behavior patterns of users in the cloud computing, understand in-depth importance of the relevant page, and improve the operating effectiveness of Web site; but because of the cloud itself has a broad distribution, it tends to have large volumes of data, and is widely distributed with heterogeneous structure, and the data is in real-time transformation, hence has greatly increased the complexity of the data structure, and brings difficulty to data mining in cloud computing. Especially in the data mining process, the constant and multiple access to databases presents new requirements to data mining in cloud computing environment.

2.3 Map / Reduce Model

Map / Reduce model in cloud computing environment is a distributed model; in dealing with a large amount of data it has been widely used [9]. Principle of Map / Reduce model is by using Map function to divide the input process into a large number of data pieces, and the data segment units will be delivered to the computer for processing; by Reduce function the processed data fragments will be integrated together, and finally output a summary; in Map / Reduce model, Master is responsible for task scheduling and sharing the data between the node; Worker is primarily responsible for task processing and data processing in cloud computing [10-11]. As shown in Figure 1

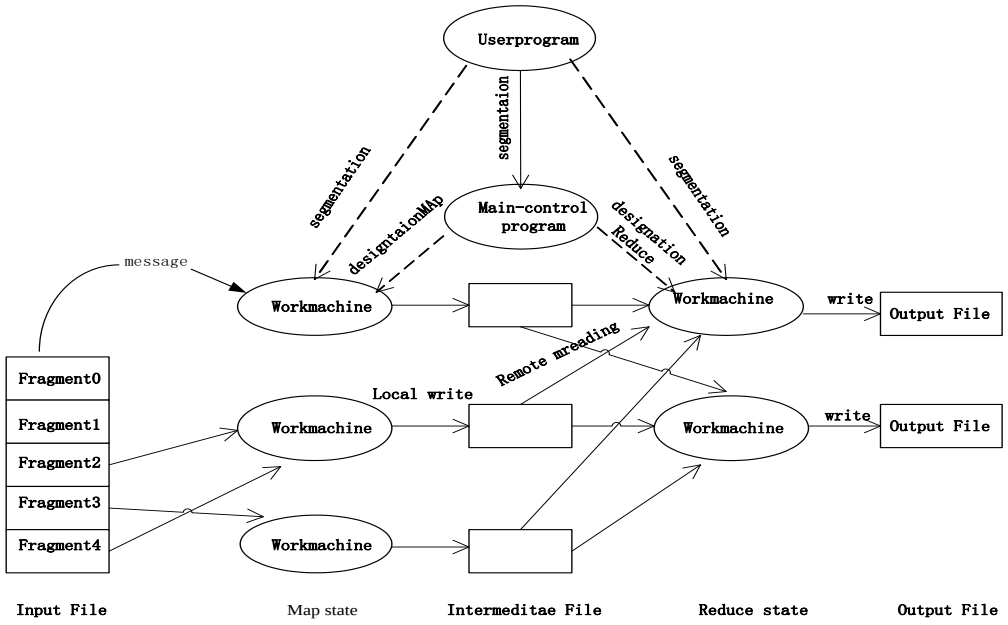

Figure 1. Map / Reduce Model

3. SVM in Cloud Computing Model Environment

3.1 Support Vector Machines with the Introduction of Penalty Factor

This paper has first introduced in the SVM penalty factor γ_i, assuming that x_i represents the input vector, F is high-dimensional inner product space; feature vector $z_i \in F$, $z_i = \sigma(x_i)$, in which σ is the characteristic function, kernel function is $k(x_i, x_l) = \sigma(x_i) \cdot \sigma(x_l)$. Then SVM optimization problems related to optimization features are:

$$\begin{aligned} &\min 1/2\|w\|^2 \\ &s.t\ w \cdot z_i + b \geq 1 - \gamma_i \quad \forall i \in I \qquad (4) \\ &\quad w \cdot z_i + b \leq -1 + \gamma_j \quad \forall j \in J \end{aligned}$$

In Equation (4), z_i means linearly separable; after the introduction of the penalty factor, related optimization problems have been transferred as:

$$\begin{aligned} &\min 1/2\|w\|^2 + \frac{c}{2}\sum_k \gamma_i^2 \\ &s.t\ w \cdot z_i + b \geq 1 - \gamma_i \quad \forall i \in I \qquad (5) \\ &\quad w \cdot z_i + b \leq -1 + \gamma_j \quad \forall j \in J \end{aligned}$$

3.2 Introduction of Distributed Support Vector Machines in Cloud Computing Model

When the data is large-scale in cloud computing system, the calculating process of SVM requires a very large amount of calculation, therefore, the data in this paper proposes using cloud computing-based distributed SVM classification algorithm for large-scale, complex cloud data, and the algorithm steps are as follows:

(1) assign data records of cloud computing environment to different computing nodes, and use the target equation of the improved support vector machines function:

$$f(x_i) = \min 1/2\|w\|^2 + \frac{c}{2}\sum_k x_i^2 \qquad (6)$$

(2) In each computing nodes, respectively, to calculate the equation value of for each node $\sum_i a_i y_i \Phi(x_i)$, $\sum_i y_i \Phi(x_i)$ 以及 $\sum_k \gamma_i^2 x_i^2$。

(3) After the x_i calculation of each node, merge summation the corresponding intermediate equations values of each computing node, and according to the Lagrange function factor vectors to obtain partial differential evaluation of three middle equations, namely:

$$\begin{cases} \dfrac{\partial f}{\partial x} = 0 \Rightarrow x = \sum_{i=1}^{n} a_i y_i \Phi(x_i) \\ \dfrac{\partial f}{\partial y} = 0 \Rightarrow y = \sum_{i=1}^{n} y_i \Phi(x_i) \\ \dfrac{\partial f}{\partial x \partial \gamma} = 0 \Rightarrow x\gamma = \sum_k \gamma_i^2 x_i^2 \end{cases} \qquad (7)$$

(4) According to (3), the expression result is 0, resulting in the value of the optimal solution x_i.

4. Mining Algorithms Introduced Support Vector Machine based on Map / Reduce Model

4.1 Algorithm Steps Description

The algorithm is distributed computing framework based on Map / Reduce model, using three selection algorithm to obtain overall frequent item sets of data; in the cloud computing environment the documentation set is the documents collection to be dealt with under the assumption in the cloud computing environment. Take $D_n = \{txt_1, doc_2, excel_3 \cdots\}$ as an example, steps are as follows:

Step 1: According to the contents of the document collection, using the Map function to convert the data block $f(D_i) \to < c_i, D_{ni} >$; through Reduce function it is expressed as $f(D_i) \to < c_i, D_{ni} >$, $< c_i, f(D_{ni}) >$

Step 2: in the output $< c_i, f(D_{ni}) >$ we introduce equation (6) to calculate object equation for each document D_{ni}; and introduce equation (7) to calculate the optimal solution set of the document: $D_{xn} = \{txt_{x1}, doc_{x2}, excel_{x3} \cdots\cdots\}$.

Step 3: take $f(D_{xn})$ in Step 2 as the second entry of Map function, and then get Map function processing data block $f(D_{xn}) \to < c_{xi}, f(D_{xn}) >$

Step 4: Repeat steps 2-3 until iterations reach 4.

4.2 Algorithm Implementation

In this paper, we use three selections algorithm in the Map / Reduce model to get overall frequent item sets, the algorithm implementation is as follows:

```
1: A variety of documents in cloud computing , Dn={txt1,doc2,excel3……},N=3
2:while(N<3)
3:{Map();
4: for(i=1;i<n;i++)
```


5: Fre(i)=Fre(i-1)+1;
6: ouput($doc_{i,}$Fre(<c_{xi},doc_{xi}>));
7: $\lambda_j=\lambda_i$. Frequency(i);
8: Endfor
9: EndMap
10:Reduce()
11:Input($doc_{i,}$Fre(<c_{xi},doc_{xi}>));
12:for(c_{xi}∈record(doc_{xi}))
13:output(c_{xi},Fre(c_{xi},doc_{xi}>));
14:endReduce;
15: $f(c_{xi}) = \min 1/2\|w\|^2 + \frac{C}{2}\sum_k x_i^2$
16: Fre(c_{xi})→c_{xi}
17:N=N+1
18:EndWhile

4.3 Algorithms Example

To further illustrate the usefulness of the proposed model in this paper, the author introduced of the concept of characteristics multi-tree from Lhe literature [12] : nodes (excluding leaf nodes) of each layer of characteristics multi-tree correspond to a collection of the feature properties; the side corresponds to different values of the data characteristics, and data characteristics is corresponding the sides of multi-tree. The result of node storage is the value of multi-tree; each node in the multi-tree saves a certain value and the number of records.

Steps are as follows:

1: Input: training samples are distributed in sub-nodes of cloud computing, labeled as X = {X1, X2 ... Xn}, reflecting support vector collection ETi of X1, X2 ... Xn; node weights in descending order, E = {E1, E2 ... En},
2: Output: SVM_N of cloud feature multi-tree
3: Content: :For k=1 to N
4: If(k==1)
5: / / from N to get all Count (N) values of E1 and the number of records, and then create a node Enodei according to statistics
6: $SVM_{n+1} = \frac{\sum_{i=1}^{n}\int SVM_n}{N}$
7:For k=1 to Count(N)
8: $ENode_{kj}$(value,Count)=CreateTree($E_{N,}$)
9:EndFo
10:else
11:// Get the statistical value of ETi that meet the conditions of each node C(N, and the number of records; create node $ENode_{kj}$ based on the results of values respectively.
12:For k=1 to Count(N)
13: For j=1 to C(N)
14: $ENode_{kj}$[Count]=CreateTree($E_{k,}$,SVM_n)
15: SVM_{N+1}=CreateTree(ET_i)

16: $SVM_{n+1} = \frac{\sum_{i=1}^{n} f(SVM_n)}{N}$
17: EndFor
18:EndFor

5. Experimental Simulation

This algorithm uses Cloudsim platform and selects network centry of the author's institute; the author has selected three supermarket near the school as cloud, and the supermarket information will be seen as a classification problem; set N given training samples $f_i = (x_i, y_i)_i^m$,in which x_i represents the input sample , y_i means the sample output; usually -1 means loss means and +1 indicates the presence. Due to retrieving books under cloud conditions related to the dynamic nature, and difficult to be captured, support vector machine adopts the kernel function $k(x, y) = \exp(-\|x - y\|^2) / \beta$, where $\beta = \frac{\sum_{i,j=1}^{m} \|x_i - y_i\|^2}{m^2}$. Extract the overall books information data and number of readers obtained from several libraries forecasts : the author extracted the main ingredient values to obtain data sets, and collected overall support vector data sets as training data, using support vector machine algorithm for global mining.

Focused on the information for each supermarket, the author extracted according to a certain proportion 1000,2000,3000 records of product information as measurement sample evenly distributed in three different databases, using Literature [13], [14] and the algorithm of this paper to compare their vector machine number, operation duration, classification accuracy rate, and the results are shown in Table 1. This paper compares SVM and distributed, centralized SVM, as shown in Table 2.

Table 1. Comparison between Distributed SVM Algorithm

Name	Algorithm	Support Vector	Run Time (sec)	Classification accuracy
Supermarket 1 Product Information	Literature[13] algorithm	70	19.34	63.27%
	Literature[14] algorithm	75	15.23	68.37%
	the proposed algorithm	85	12.05	80.15%
Supermarket 2 Product Information	Literature[13] algorithm	55	36.12	49.18%
	Literature[14] algorithm	62	32.14	53.27%
	the proposed algorithm	67	28.96	80.19%
Supermarket 3Product Information	Literature[13] algorithm	80	56.21	70.21%
	Literature[14] algorithm	84	39.72	73.19%
	the proposed algorithm	87	20.23	85.27%

From Table 1 it is found that, when compared the proposed algorithm with Literature [13],

[14] algorithms, it has saved the running time, but the classification accuracy is moderate, showing that the algorithm in this paper is effective.

Table 2. Comparison of the Proposed SVM, Distributed SVM and Centralized SVM

Record number of test information	Support Vector Machine	Support Vector	Time (sec)	Accuracy (%)
1000	the proposed SVM	152	27.15	89.23
	centralized SVM	184	20.34	79.14
	distributed SVM	194	17.47	69.15
2000	the proposed SVM	915	524.29	89.25
	centralized SVM	1232	429.24	82.35
	distributed SVM	3128	352.26	79.25
3000	the proposed SVM	7231	1562.34	88.29
	centralized SVM	8251	1320.29	80.18
	distributed SVM	9217	10256.16	75.26

From Table 2 it can be found that when the data is in distributed environment, the proposed SVM by introducing penalty factor only has to summarize data set and supports vector properties characteristic values for aggregation; the amount of data it needs is far less than the amount of distribution record data. Meanwhile, SVM herein has higher accuracy than the centralized and distributed SVM. From the overall performance point of view, by conducting book information mining in SVM environment proposed by this paper has more advantages than that of the centralized and distributed SVM in real environment.

6. Conclusion

How to dig out the useful information in the cloud computing environment is the focus of current research, but because the data in cloud computing itself is dynamic with uneven distribution, the author has proposed to introduce a penalty factor in SVM while in the cloud computing model Map / Reduce to adopt three options algorithm , which to a certain extent has improved the accuracy of the information mining under cloud computing environment; the experiment proved that the proposed mining algorithms could achieve good results in cloud computing mining resources.

References

[1] C. Miao, “Web data mining based on cloud computing”, Computer Science, vol. 28, no. 10, **(2011)**, pp. 146-149.
[2] Y. Yi, “Data mining based on cloud computing”, Microelectronics and Computer, vol. 30, no. 2, **(2013)**, pp. 161-164.
[3] D. Jing, “Data mining service model in cloud computing environment”, Computer Science, vol. 39, no.

6, **(2012)**, pp. 217-219.
[4] Z. Youlin, "Analysis of cloud services in data mining", Intelligence theory and practice, vol. 35, no. 9, **(2012)**, pp. 33-36.
[5] G. Xin, "A data mining algorithms in dynamic cloud model", Mini-Micro Systems, vol. 34, no. 12, **(2013)**, pp. 2749 -2754.
[6] Y. Lai, "Parallel data mining method based on Hadoop cloud platform", vol. 25, no. 5, **(2013)**, pp. 934-944.
[7] W. Xiangxi, "Assessment of web health status based on support vector machine and the cloud model", Beijing University of Posts and Telecommunications, vol. 35, no. 1, **(2012)**, pp. 10-14.
[8] L. Xuezhi, "Classification predict application of distributed support vector machine under cloud computing platform in the coal industry", Coal technology, vol. 32, no. 11, **(2013)**, pp. 248-250.
[9] Li Zhen, "Improved Map-Reduce model in cloud computing environment", Computer Engineering and Applications, vol. 38, no. 1, **(2012)**, pp. 27-30.
[10] L. Jing, "Map-Reduce job scheduling algorithm based on improved leapfrog strategy", Computer Applications Research, vol. 30, no. 7, **(2013)**, pp. 1999-2002.
[11] D. Linlin, "Massive data efficient Skyline query processing based on Map-Reduce", Journal of Computers, vol. 34, no. 10, **(2011)**, pp. 1785-1795.
[12] C. Zhenzhou, "Multi-tree expression of curve", Surveying and Mapping, vol. 42, no. 4, **(2013)**, pp. 602 - 606
[13] Q. Yihui, "Telecommunications industry customer churn prediction based on random forests and single class support vector machines", Xiamen University (Natural Science Edition), vol. 52, no. 5, **(2013)**, pp. 603-605.
[14] Q. Tao, "Improved support vector machine applications in the telecommunications customer loss forecasts", Computer Simulation, vol. 28, no. 7, **(2011)**, pp. 329-332.

Authors

Lu Shuhong (1980.2-), female, senior lecturer, research team leader, graduate; research direction: software technology, network security.

论文题目	一种改进的SVM在云计算信息挖掘中的研究		
所属主题	其他主题		
第一作者			
姓名	吕树红	职称/学位	高级讲师
单位	正德职业技术学院	邮编	211106
地址	南京市江宁经济开发区将军大道18号		
电话		手机	18057525807
Email	1713754023@qq.com		
第二作者			
姓名		职称/学位	
单位		邮编	
地址			
电话		手机	
Email			

