Research on the Massive Data Classification Method in Large Scale Computer Information Management

huangyun

Chongqing electronic engineering of Career Academy, Chongqing 471331, China

Keywords: large scale computer information management; massive data; Bayesian belief network; the tacit understanding

Abstract. In the process of the massive data classification in large-scale computer information management system, due to the large amount of data, and included large number of feature data, the correlation of data is reduced, resulting in low efficiency of computer operation. A model for massive data depth classification mining based on belief network is put forward. According to the relation between the probabilities of data in all the data domain, the correlation between knowledge and data domain can be inferred. Through the training sample set find the most suitable Bayesian belief network for the sample data, then according to the possible management structure and the tacit understanding degree between data samples, the optimal solution within data management structure of large scale computer. The experimental results show that, using the improved algorithm for massive data classification processing, can improve the accuracy of classification, and achieve satisfactory results.

Introduction

Data classification is the core part of computer data extraction [1]. By using the method of data classification, it can make reasonable classification of the massive data according to the user's needs, so as to realize the accurate extraction of the effective network data, improve the operation efficiency of network [2], the method has a high value of application in the field of database fields, becoming the hot topic in required research field of data management in large computer [3]. Currently, data classification algorithm mainly includes the methods based on the support vector machine, wavelet transform and the data gain algorithm [4-6]. Among them, the most commonly used is the wavelet transform algorithm.

In order to avoid the defects of traditional methods, a model for massive data depth classification mining based on belief network is put forward. According to the relation between the probability of data in all the data domain and possible data management structure and the tacit understanding between samples, the optimal classification data within the data management structure in the large computer is obtained. The experimental results show that, using the improved algorithm for massive data classification processing, can improve the accuracy of classification, and achieve satisfactory results.

The principle analysis of massive network data classification

Massive network data classification is a powerful guarantee of network communication. Usually, the distribution of big data in network data is the normal distribution, which can be described by $n \sim N(0,\delta^2)$. Using the following formula calculate the transformation parameters of large data in network:

$$Z = Y + N(0,\delta^2) \quad (1)$$

The network data are treated by using wavelet transform, the data is still subject to the normal distribution, which can be described using the following formula:

$$Z = Xz = Xy + N(0,\sigma^2) = Y + N(0,\sigma^2) \quad (2)$$

It can be seen that the possibility of network data subjected to normal distribution in the interval $(\nu - 3\sigma, \nu + 3\sigma)$ is 0.9991. so, the data of information management in the large scale computer

can be divided into two sections according to wavelet transform parameters. Assuming that $\left|e_k^l\right| > 3\sigma$, the degree of importance of this data is higher, otherwise, it is lower. Among them, e_k^l is the wavelet transform parameter included data.

If $3\sigma \geq U$, $\left|e_k^l\right| \geq 3\sigma \geq U$ and $\left|e_k^l\right| U \geq 3\sigma U \geq U^2$, $U \geq \frac{U^2}{\left|e_k^l\right|}$ can be obtained, and finally, the threshold $\frac{U^2}{\left|e_k^l\right|}$ of wavelet transform processing is acquired. In the above mentioned status, using the following formula can make wavelet transform processing:

$$\hat{e}_k^l = \begin{cases} \operatorname{sgn}(e_k^l)\left(\left|e_k^l\right| - \frac{U^2}{\left|e_k^l\right|}\right), \left|e_k^l\right| \geq 3\sigma \\ \operatorname{sgn}(e_k^l)\left(\left|e_k^l\right| - U\right), \quad U \leq \left|e_k^l\right| < 3\sigma \\ 0 \qquad\qquad\qquad , \quad \left|e_k^l\right| < U \end{cases} \tag{3}$$

when $3\sigma < U$, $\left|e_k^l\right| \geq 3\sigma$ and $\left|e_k^l\right| 3\sigma \geq 3\sigma^2$, $U \geq \sigma \geq \frac{(3\sigma)^2}{\left|e_k^l\right|}$ can be obtained, and finally, the threshold $U \geq \sigma \geq \frac{(3\sigma)^2}{\left|e_k^l\right|}$ of wavelet transform processing is acquired.In the above mentioned status, using the following formula can make wavelet transform processing:

$$\hat{e}_k^l = \begin{cases} \operatorname{sgn}(e_k^l)\left(\left|e_k^l\right| - \frac{(3\sigma)^2}{\left|e_k^l\right|}\right), \left|e_k^l\right| \geq 3\sigma \\ 0 \qquad\qquad\qquad , \quad \left|e_k^l\right| < 3\sigma \end{cases} \tag{4}$$

The model for depth classification mining of massive data based on Bayesian belief network

Bayesian belief network.

Definition 1:setting a arbitrarily variable set $x = \{x_1, x_{2\leq} \cdots, x_n\}$, if a combination contingent probability on x is disperse, then it can be defined using Bayesian belief network as follows:

$$B = \langle G, \theta \rangle \tag{5}$$

Among this, X is a m dimensional vector quantity; G is a directed acyclic graph, the arcs represent a functional dependence; θ represents a group of parameter to make quantization network.

Definition 2: If there is an arc from the variable Y to X, then Y is X's parents or the direct precursor, while X is the successor of Y. Once given its parents, all variables in acyclic graph in the graph independent of the non-successor of nodes, all variables of parents of X_1 in G are as a set $P_a(X_1)$.

The massive data mining process based on Bayesian belief network. Setting a group of data training sample $D = \{x_1, x_2, \cdots, x_n\}$, x_1 is X's example, through estimation function $S(B|D)$ find a Bayesian belief network which is suitable for this sample. Using function $S(B|D)$ can estimate the tacit understanding degree between all the possible management structures and data samples, to the optimal classification data.

Bayesian belief network algorithms are described as follows:

Input: massive data training sample set $D = \{x^1, x^2, \cdots, x^n\}$,initialize network B_0, and evaluate the function $S(B|D) = \sum_1 S(X_1|Pa(X_1), D)$, parameter k

Output:the optimal network from $1, 2, \cdots,$ to n

(1) compress: according to D and B_{n-1}, the use of candidate compression, from X_1, X_2 ,…, X_n, select a candidate farther set $C_i^n(|C_i^n| \leq k)$ for X_1. here, a digraph $H_n = (x, E)$ is defined, wherein, $E = \{X_j \to X_i | \forall i, j, X_j \in C_i^n\}$.

(2) maximization: find a Bayesian belief network $B_n = \langle G_n, \Theta_n \rangle$ which can maximized evaluate the

function $S(B_n|D)$, wherein, $G_n \subset H_n, \forall X_1, Pa^{C_n}(X_1) \subseteq C_i^n$ and return B_n.

Through analyzing evaluation function, and applying candidate compress algorithm, it can complete the filter that data X_i will become the data set $Pa(X_1)$. Select the most probability variables in k can become X_i, to get the correlations density between variables, it needs to integrate into dependence function $I(X,Y)$:

$$I(X,Y) = D_{KL}(P(X,Y)|P(X)P(Y) \tag{6}$$

Among them, $D_{KL}(P(X)\left|Q(X) = \sum_X P(X)\log\frac{P(X)}{P(Y)}\right.$.

The candidate compress algorithm is as following:

Input: the weight calculation of a Bayesian belief network B_n $S(B|D)$ in data set $D = \{x^1, x^2, \cdots, x^n\}$, parameter k

Output:for all data X_i, return a k candidate C_i set for all of X_i, $i = 1,2,\cdots,n$.

(1) for all X_j ,calculate $I(X_i, X_j), X_j \neq X_i$ and $X, \notin Pa(X_i)$

(2)select the element which has the highest weight $k-1$, $l = |Pa(X_1)|$ candidate set, $C_i = Pa(X_1)Y\{x1,\cdots,xk-1\}$ back to $\{C_i\}$.

Experimental results and analysis

By using simulation software matlab 7.1 construct the experiment environment. The number of all the data in a database of large computer is set up as 1000. The number of feature data types needed for classification are 20 kinds. From the above database, 10 different characteristics of the data are randomly selected, the specific situation can be described with the figure 1. In the figure, each color represents one kind of the characteristics of data.

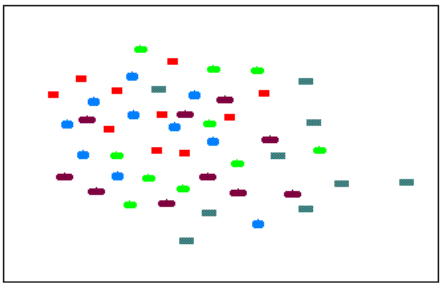

Figure 1 the distribution diagram of different attribute data

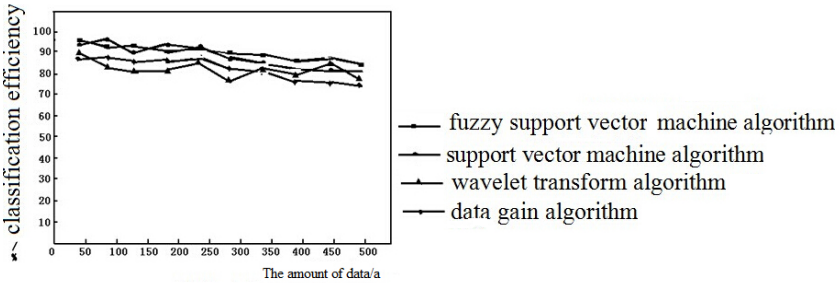

Figure 2 the classification results when the types of feature data are few

According to the experimental results in figure 2, it can be learned, if the types of feature data is relatively few, the efficiency of data classification utilizing the improved algorithm is similar to and the traditional algorithm.

When the type of the feature data in the database are more, using different algorithms for feature data classification respectively, the feature data classification results obtained can be used to describe by figure 3:

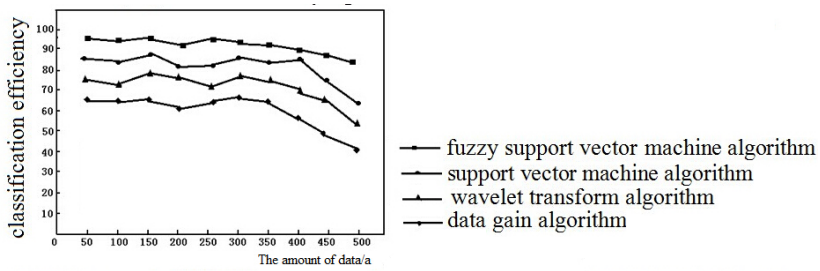

Figure 3 the classification results when the types of feature data are more

According to the experimental results in figure 3, it can be learned, if the types of feature data is relatively more, the efficiency of data classification utilizing the improved algorithm is higher than the traditional algorithm.

When the type of the feature data in the database are fewer, using different algorithms make 5 times redundant data classification, the feature data classification results obtained can be used to describe by table 1:

Table 1 Different algorithms data table data when the types of feature data are fewer

The number of experiments	Accuracy of classification using support vector machine algorithm（%）	Accuracy of classification using wavelet transform algorithm （%）	Accuracy of classification using data gain algorithm （%）	Accuracy of classification using fuzzy support vector machine algorithm（%）
1	96	95	94	97
2	97	92	92	98
3	95	91	89	97
4	94	94	91	96
5	95	92	93	95

In circumstances of more types of feature data, the use of different algorithms to make 5 feature data classification, the data in the experimental process are analyzed, to get the experimental results in table 2:

Table 2 Different algorithms data table data when the types of feature data are more

The number of experiments	Accuracy of classification using support vector machine algorithm （%）	Accuracy of classification using wavelet transform algorithm （%）	Accuracy of classification using data gain algorithm （%）	Accuracy of classification using fuzzy support vector machine algorithm（%）
1	89	78	69	98
2	87	72	67	98
3	92	78	67	98
4	85	77	65	96
5	78	58	42	92

Through the above experiments it can be learned, using the improved algorithm for massive data classification in large-scale computer information management, can avoid the defect of the poor correlation between the mass data in the database, so as to improve the efficiency of feature data classification.

Conclusions

For the problem of in the process of the massive data classification in large-scale computer information management system, due to the inevitable low correlation of data caused by large amount of data, resulting in low efficiency of computer operation, a model for massive data depth classification mining based on belief network is put forward. According to the relation between the probabilities of data in all the data domain, the correlation between knowledge and data domain can be inferred. Then, through the training sample set find the most suitable Bayesian belief network for the sample data, then according to the possible management structure and the tacit understanding degree between data samples, the optimal classification data within data management structure of large scale computer. The experimental results show that, using the improved algorithm for massive data classification processing, can improve the accuracy of classification, and achieve satisfactory results.

References

[1] Jiang Heng, Chang Jianping. Study of SAR dynamic target detection of based on multiple signal classification method [J]. Computer simulation 2011.9:264-267.

[2] Wan Changxuan, Liu Xiping. XML database technology. Beijing: Tsinghua University press, 2008:137-200

[4] Sun Yizhong. The theory and basis of application of XML. Beijing: Beijing University of Posts and Telecommunications press.2000:130-134

[3] Zhou Xusheng, Li Shuang. Modeling and Simulation of automatic web page classification [J]. Computer simulation, 2011.10:121-124.

[4] Han Tong. Track and manage the number of concurrent users to improve the efficiency of database system [J]. Information technology and informatization, 2011.5:49-50.

[5] Longqian, Guo Jinchi. The investigation and analysis of 985 University's Library self-built database. [J]. Researches in library science, 2010.18:27-31.

[6] Li Wenjie. Optimal design scheme of large database ORACLE database [J]. Technology wind, 2011.19:145.