Application of the massive data precision classification in e-commerce based on big data

Li Ting

Jincheng college of sichuan university, Sichuan, Chengdu, 611731

Keywords: big data; data classification; recommendation system; relational decision-making tree

Abstract. Accurate classification of massive data in the field of electronic commerce, can service consumers to the greatest degree. The variety and quantity of goods fast growth make a large number of useful products information would be submerged in massive data. The traditional electronic commerce data classification system applied in the user data validation, is affected by massive information, to make data validation process time-consuming and low efficiency. The massive data precision classification system of electronic commerce based on big data is proposed. Large amounts of data is made big data relational decision-making calculation, to get the relevance between all data. Based on the above data relevance, the relational decision-making tree is established, to obtain user's interest data recommended goal. The experimental results show that, by using the improved algorithm for the user massive interest data classification in electronic commerce, can effectively reduce the time to confirm the user interest commodity information data, ensure data quality to meet the requirement of electronic commerce.

Introduction

With the rapid increase of data classification technology, the technology has been widely used in database management in various industries, playing a more and more important role [1]. In the field of electronic commerce, with the rapid increase of the number of and types of goods and commodities, lots of useful information will be submerged in the vast amount of data, therefore, how to make efficient commodity recommendation based on user interest [2], has become the core problem of classification recommendation system needed to study in electronic commerce [3]. In the condition of massive commodity information data, user interest data classification recommendation system, become the [4] key subject to study in the field of electronic commerce. Currently, the precision massive data classification system in electronic commerce mainly includes the system based on the improved support vector machine algorithm [5], the Gauss model and fuzzy clustering algorithm. One of the most commonly used is the accurate massive data classification system in electronic commerce field based on improved support vector machine algorithm. Due to the accurate massive data classification system in electronic commerce field plays an irreplaceable role in the commercial areas, it has a broad prospect for development, and has become the focus of many experts research topic.

Principle analysis of accurate massive data classification in electronic commerce field

The accurate classification of massive data in the field of electronic commerce, can service consumers to the greatest degree. Usually, the distribution of commodity information data in electronic commerce field follows normal distribution, which can be used $V = \{v_1, v_2, \cdots, v_q\}$ to describe. Using the following formula can calculate the transformation parameters:

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

The above commodity data are made wavelet transform processing, the data are still subject to the normal distribution, which can use formula (2) to describe:

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

The probability of the commodity information data subjected to normal distribution in the interval $K = \{k_1, k_2, \cdots, k_p\}$ is 0.9991, therefore, it can be according to the wavelet transform parameters

to divide the massive commodity information data in electronic commerce into two different parts. If $L_p(Y,B)=\sum_{k=1}^{q}\sum_{l=1}^{e}y_l^p f_{kl}^2(z_k,b_l)$, the degree of importance of the data is relatively high, on the contrary, its importance is relatively low. Wherein, $e,q,p,d=1$ is the wavelet transform parameter which includes data.

Use the following formula can deal with wavelet transform:

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

Accurate massive data classification method based on big data in e-commerce

Using the traditional algorithm for accurate massive data classification in electronic commerce, commodity data classification system in user's interest confirmation, is vulnerable to be affected by massive information, causing a time-consuming process of interest data validation and low efficiency. Therefore, the precision massive data classification system based on big data in electronic commerce is proposed.

Clustering processing of association big data. Setting that the composed sequences of massive data in e-commerce can be described by $V=\{v_1,v_2,\cdots,v_q\}$, among them, v_l is used to describe the l-th commodity information data in this sequence, the corresponding property of the above data can be described by $K=\{k_1,k_2,\cdots,k_p\}$. Using fuzzy clustering method, all commodity information data can be classified. Specific methods are as follows:

Using the following formula describe the fuzzy clustering objective of massive data:

$$L_p(Y,B)=\sum_{k=1}^{q}\sum_{l=1}^{e}y_l^p f_{kl}^2(z_k,b_l) \tag{4}$$

Setting that the state parameters of massive data in electronic commerce can be described by $e,q,p,d=1$, the clustering center can be described by $B_{(d)}=(b_1,b_2,\cdots,b_e)$, using the following formula make update processing of all the fuzzy clustering center:

$$\begin{aligned} &y_{kl}=\frac{1}{\sum_{m=1}^{e}\left[\dfrac{f_{kl}}{f_{km}}\right]^{\frac{2}{p-1}}} \text{ 当 } f_{kl}\neq 1 \\ &y_{kl}=0 \text{ if } e_{kl}=0, l\neq m \\ &y_{kl}=1 \text{ if } e_{kl}=0\end{aligned} \tag{5}$$

Using the following formula calculate the average value of massive commodity information data:

$$b_l=\frac{\sum_{k=1}^{q}y_{kl}}{\sum_{k=1}^{q}y_{lm}} \tag{6}$$

b_d and $b_{(d+1)}$ are compared, assuming that the data conforms to the following constraints, it can realize the fuzzy clustering processing of massive data in the electronic commerce:

$$\left|b_d-b_{(d+1)}\right|\leq\varphi \tag{7}$$

In the environment of massive data in electronic commerce, the range of objective function value of fuzzy clustering is shrinking. In the fuzzy clustering process of massive data, it can effectively avoid the disadvantages of local minimum in the iterative process, and access to massive data clustering structure.

The establishment of the relational decision-making tree for user interest data in e-commerce. The set composed by massive data is $Z=\{(z_k,a_k)\,|\,k=1,2,\cdots,total\}$, the k-th information data in the set can be described by $z_k=(z_{k1},z_{k2},\cdots,z_{kf})$. According to the following formula, it can calculate the value of expectations of interest data in mass goods information:

$$K(q_1,q_2,\cdots,q_p)=-\sum_{l=1}^{p}\frac{q_l}{total}\log_2(\frac{q_l}{total}) \tag{8}$$

In the set composed by massive data in electronic commerce, all the set consisted by attributes of data is $C_h(h=1,2,\cdots,f)$, use the following formula decompose the data attribute:

$$G(C_h)=\sum_{u=1}^{s}\frac{q_{1u}+\cdots+q_{pu}}{total}K(q_{1u},q_{2u},\cdots,q_{pu}) \tag{9}$$

Using the following formula build the decision-making tree of interest data recommendation in massive commodity information:

$$C_k=MIN+\frac{MAX-MIN}{Q}\times k \tag{10}$$

in above formula， $k=1,2,\cdots,Q$.

According to the method of above described, the acquired decision-making tree structure of massive data can be used to describe as below :

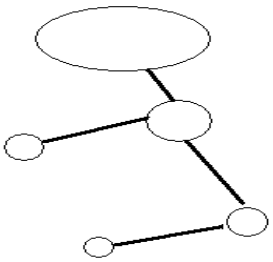

Fig. 1 decision-making tree structure of the commodity information data

The interest information data which ratio of gains reaches to maximum value under massive data in electronic commerce are taken as a branch of the decision-making tree, to build the decision-making tree of interest commodity information data under massive data environment.

Experiment results and analysis

In order to verify the effectiveness of improved algorithm, it needs an experiment. The experimental environment is Visual C++6.0. The amount of the sample data in electronic commerce commodity information database is 1000, the type of all the merchandise information data is 15, in which randomly select 5 commodity information attributes as the experimental object, the number of all commodity information data is P, the number of species of all commodity information data is p, all the massive data set is composed of $\{b_1,b_2,\cdots,b_p\}$, all the data set consisted by the mass data attribute is $\{c_1,c_2,\cdots,c_p\}$, commodity information data b_j belongs to the probability λ of attribute c_k.

Using the following formula calculate the user's interest data confirmation time in the environment of massive data in electronic commerce:

$$\varphi=\frac{\sqrt{b_j-b_{j-1}^2}}{|c_k-1|} \tag{11}$$

The cost of time of that user confirms interest data, is an important index to measure different data classification methods. In the state of lower sample complexity, respectively, using the traditional algorithm and the improved algorithm make the customer interested commodity information classification, the time needed to expend of user classification recognition could be used to describe as below:

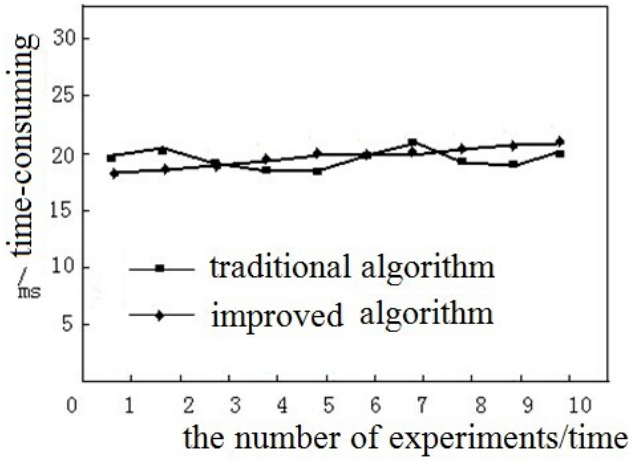

Figure 2 time-consuming of user classification confirmation when sample complexity is lower

According to the above it can be seen that, when the sample complexity is lower, the use of improved algorithm for user information data classification, and the time spent by user to confirm are basically the same with traditional algorithm.

All the sample data are as the experimental object, make use of the traditional algorithm and the improved algorithm for the above commodity information data to make user information data classification, the needed spend time of user to confirm can use figure 3 to describe:

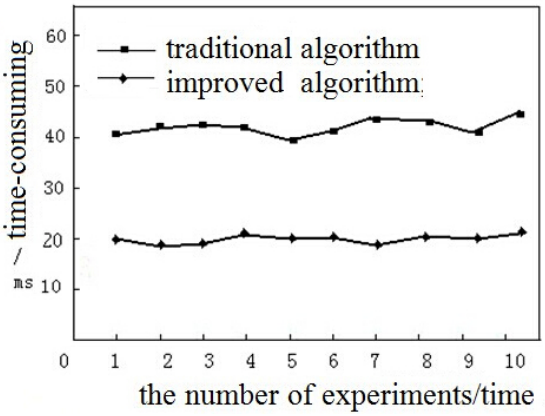

Figure 3 time-consuming of user confirmation classification when sample complexity is high

According to the above it can be seen that, when the sample complexity is higher, the use of traditional algorithm for user information data classification, and the time spent by user to confirm is about 17ms, while the use of improved algorithm for user information data classification, the time spent by user to confirm is about 9ms, which fully embody the advantages of the improved algorithm used in higher sample complexity.

The results of user information data classification can be used to describe as table 1:

Table 1 experimental results of low sample complexity

The number of experiments	Time consuming of traditional algorithm（ms）	Time consuming of improved algorithm（ms）
1	18	17
2	17	19
3	20	18
4	21	17
5	19	16
6	18	18
7	20	20
8	21	18
9	19	19
10	20	21

In status of high sample complexity, the use of the traditional algorithm and the improved algorithm classify the data information, the user's confirmation results can be used to describe as table 2:

Table 2 experimental results of high sample complexity

The number of experiments	Time consuming of traditional algorithm（ms）	Time consuming of improved algorithm（ms）
1	41	20
2	42	19
3	43	19
4	43	21
5	39	20
6	40	20
7	42	19
8	41	20
9	40	20
10	43	21

According to the experiment on the table 2, it can be learned, that using improved algorithm for, customer interest commodity information data classification in the massive commodity information data environment, can avoid the traditional algorithm's defect of poor convergence due to the excessive complexity of massive data, so as to improve the accuracy rate of commodity information classification.

Conclusions

Aiming at the problems of that traditional electronic commerce data classification system applied in the user data validation, is affected by massive information, to make data validation process time-consuming and low efficiency, the massive data precision classification system of electronic commerce based on big data is proposed. Large amounts of data is made big data relational decision-making calculation, to get the relevance between all data. Based on the above data relevance, the relational decision-making tree is established, to obtain user's interest data recommended goal. The experimental results show that, by using the improved algorithm for the user massive interest data classification in electronic commerce, can effectively reduce the time to confirm the user interest commodity information data, ensure data quality to meet the requirement of electronic commerce.

References

[1] Guo Luodan, Kong Jinsheng. Application of MRBF neural network in image classification [J]. Computer engineering and design, 2009, 30 (13): 3154-3156.

[2] Deng Huawu, Clausi D A. Gaussian MRF rotation-invariant features for image classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(7): 162-165.

[3] Zhang Jianping, Liu Xiya. Research and application of K means algorithm based on clustering analysis [J]. Computer research and application, 2007, 24 (5): 166-168

[4] Xie Wenlan, Shi Yuexiang, Xiao Ping. Application of BP neural network in the natural image classification [J]. Computer engineering and application, 2010, 46 (2): 162-166.

[5] Zhang Jinshui, He Chunyang, Pan Yaozhong, et al. Study on classification of high spatial resolution remote sensing data with multi-source information composition based on SVM [J]. Remote sensing journal, 2006, 10 (1): 49-57.