Research on Big Data Classification Based on K-Means Bayes Algorithm in Cloud Storage Environment

Changhong Wu

Guangdong Preschool Normal College in Maoming (China Guangdong525200)

Keywords: Cloud Storage Environment; K-Means; Bayes; European Distance; Missing Value

Abstract: Cloud storage environment provides a broader space for the storage and extraction of massive big data, but also puts forward higher requirements for the performance of data classification algorithm. In view of the shortcomings of the accuracy and poor convergence performance of existing big data classification methods, this paper proposes the research of big data classification method based on K-means Bayes algorithm. Using Bayes theory to calculate the posterior probability value of data set, using k-means algorithm to improve the generalization performance of Bayes model in small sample environment; selecting cluster center and calculating the Euclidean distance from current data set to cluster center, extracting the missing value index of data set, and realizing the accurate classification of target big data set. The simulation results show that the accuracy of the proposed method is due to the traditional classification method, and the convergence performance of the algorithm model is better.

1 Introduction

Driven by the Internet and the information industry, big data technologies have been integrated into all aspects of people's lives [1]. The cloud computing framework system is one of the main tools for analyzing and processing big data [2-3]. As the core of the next-generation data processing technology, cloud computing is an extension and extension of parallel computing and distributed computing. On the Hadoop cloud resource platform [4], users can flexibly make use of cloud computing and cloud storage functions to realize the allocation processing and classification management of big data resources. Under the big data environment, the amount of data is huge and the structure is complex. Big data stored in the cloud is not only easier to retrieve and use the data, but also to ensure the security of the data [5]. Big data cloud storage is based on distributed computing, which reduces the reliance on a single local server. The efficiency of data storage and reading has improved significantly. With the expansion of the total size of big data and the growth of user groups, cloud storage will Play a more critical role. The big data system contains a large amount of semi-structured and unstructured data [6], such as video files, picture files, audio files, etc. Before these data are implemented in the cloud, they should be classified and processed based on specific algorithms to ensure data Integrity and reliability. In the cloud storage process, the existing big data classification algorithms, such as the intelligent SVM algorithm and deep learning classification algorithm, generally suffer from the disadvantages of large data classification processing overhead, low accuracy, and high classification error. Means Bayes theory classification algorithm research, combining the advantages of traditional Bayesian theory in data classification and generalization processing with k-means clustering algorithm, can maximize the preservation of detailed feature information in the original large data set, improve Efficiency and accuracy of big data classification processing and cloud storage.

2. Bayes' theorem and algorithm description

In reality, big data resources are collected in various types, different sizes, and include a large amount of semi-structured and unstructured data, so they cannot be used directly. The Bayes

classification algorithm is a classic statistical data classification scheme. The algorithm assumes that the attribute values of the current data are directly independent of other attribute values. The algorithm provides an inference method to obtain the best decision results through observation of existing data. Let be a certain hypothetical condition of the model and the label of the unknown data sample. As far as the classification of big data is concerned, the probability of the hypothetical condition appearing under the sample condition is expressed as, this type of probability is the posterior probability of the hypothetical condition, and To assume the prior probability of the condition, the theorem of the Bayes classification algorithm can be described as:

$$p(H/X)=\frac{p(X/H)\,p(H)}{p(X)} \tag{1}$$

The Bayes algorithm uses the original sample information and historical prior information to determine the probability of subsequent events, and writes the data sample as the sum of a dimensional vector:

$$X=\{x_1,x_2,.....,x_n\} \tag{2}$$

Assume that the target sample has a category. At this time, the posterior probability of the data sample is predicted according to the given data category. When the conditions $p(c_n/X)>p(c_i/X),n>i$ are met, the Bayes algorithm theorem can be known:

$$p(c_n/X)=\frac{p(X/c_n)\,p(c_n)}{p(X)} \tag{3}$$

In order to further classify unknown big data samples, first calculate the value of any sum in the set of categories, and then further allocate the data samples to the samples of the instruction according to the specific sample rules. However, the naive Bayes algorithm will reduce the classification performance of the model when the data is missing or incomplete. Therefore, this paper optimizes the classic Bayes algorithm based on the k-means algorithm to improve the data classification and clustering ability of the algorithm in small sample conditions.

3. Performance optimization based on k-means algorithm

Suppose the collection of data samples that need to be classified is represented as:

$$S=\{X_1,X_2,...,X_j,...,X_k\} \tag{4}$$

The k-means algorithm calculates the Euclidean distance between the set of samples and determines the degree of clustering of the set to optimize the classic Bayesian algorithm. The calculation process is as follows:

$$d(X_q,X_j)=\sqrt{(x_{q1}-x_{j1})^2+(x_{q2}-x_{j2})^2+...+(x_{qk}-x_{jk})^2} \tag{5}$$

Then calculate each data object in the sample set. The input of the k-means algorithm model is the data set, and the input is the convergence clustering effect of the compound Euclidean distance. In the initial stage of clustering, data objects are randomly selected as the central point in the collection. The Euclidean distance determination method based on formula (5) determines the distances of each set to the cluster, respectively. After recalculating the cluster center position under the Bayes classification algorithm, adjusting the original score of the original data object is a classic Bayes algorithm model in a small sample Under the conditions, accurate classification of the original samples can also be achieved.

4. Research on big data classification processing based on k-means Bayes under cloud storage

The current cloud computing management framework represented by Hadoop also has system vulnerabilities. The theft and loss of user data occurs frequently. Therefore, it can be seen that the

secure classification management of original big data is still necessary in the cloud storage environment. In the cloud environment, in order to improve the efficiency of big data classification and aggregation, multiple servers are often used at the same time. Even after the original data is pre-processed, malicious code may still exist, especially for foreign trade transactions of multinational companies. The problem of security classification management of data is particularly prominent. The k-means Bayes data classification management scheme proposed in this paper combines the advantages and advantages of the classic Naive Bayes method in processing uncertain events and the advantages of the k-means method in data clustering processing to solve the big data classification accuracy of the Naive Bayes method. Low deficiency. The implementation steps of data classification method based on k-means Bayes in cloud storage environment are as follows:

1: Determine the attribute value of the original big data sample and the number of classification clusters, and label each sample.

2: Divide all the pre-processed sample sets into two subsets with certain correlation, and record the specific element composition of each subset.

3: Determine the category of the target sample, use the Bayes method to classify the big data samples according to the posterior probability, and build a complete subset of big data classification attributes.

4: Use k-means algorithm to deeply optimize the naive Bayes scheme, determine the location of the cluster center, and perform a cluster analysis on the complete large data subset to identify the Euclidean distance between the data to be classified and the cluster center.

5: Calculate the similarity level of various types of data in the complete subset according to the characteristics of the missing data subset, and perform corresponding processing according to the degree of data loss to obtain missing values:

$$\eta_i = \sum_{i=1} p\left(c_i / X\right) \times \log \frac{p\left(x_i, x_n \middle| c_i\right)}{p\left(x_i \middle| c_i\right) p\left(x_n \middle| c_i\right)} \tag{6}$$

6: Based on the k-means missing data clustering method, the Bayes model is hierarchically transformed, so that the Bayes model has the characteristics of a two-layer structure; the distance between the current position of the cluster center and the original data attributes is measured The correlation and interdependence among them, the original data set is divided into a strong attribute set and a weak attribute set, and the posterior probability of each data set is calculated separately to obtain the optimal classification result. After processing the k-means Bayes method in the large data set, the original redundant data, interference data, malicious code, etc. are removed, which guarantees the integrity of the details and characteristics of the original big data to the greatest extent, especially for ensuring the information of cloud storage users Security issues are of great value.

5. Experiment and simulation

In order to verify the practical application effect of the big data classification method proposed in the article, a simulation environment was established under the Hadoop framework. The relevant characteristics of the selected test dataset are shown in Table 1

Table 1. Large data sets for experiments

NO	DATA SET	number	Attributes
1	Iris	350	3
2	Seeds	220	4
3	Wine	270	8

The Hadoop network framework system consists of 1 master node and 5 slave nodes. The basic hardware and software configuration is as follows:

Table 2. Software and hardware parameter configuration of the simulation environment

Masternode			Slave node		
1	CPU	coreI7	1	CPU	coreI5
2	Frequency	3.6Ghz	2	Frequency	3.2Ghz
3	RAM	16	3	RAM	8
4	ROM	2T	4	ROM	1T
5	system	Linux	5	system	Linux

First, the classification accuracy of the proposed classification method on the Iris, Seeds, and Wine datasets is compared. In order to make the comparison of the results more intuitive, the traditional SVM method and deep learning method are introduced to participate in the comparison. The statistical results are as follows:

Table 3. Comparison of classification accuracy of large data sets

data set	Classification method accuracy%		
	k-means Bayes	SVM classification method	Deep learning classification method
Iris	99.45	93.32	88.15
Seeds	98.75	94.52	91.17
Wine	99.13	91.17	92.92

From the statistical data results, it can be known that the k-means Bayes classification method proposed in this paper has better classification accuracy for the three different types of data sets, with an average value exceeding 99%. Compared with the method in this paper, the classification accuracy rate has a big gap. Among them, the SVM method has a poor classification effect on the Wine dataset, and the deep learning classification method is not sensitive on the Iris dataset with a weak data structure. Second, compare and analyze the convergence performance of three different classification methods, as shown in the following figure:

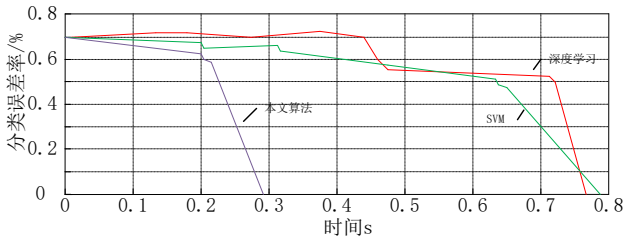

Figure 1. Comparison of classification error rate convergence performance of big data classification methods

The results of the convergence performance curve of the big data classification error rate show that the traditional SVM classification method and the deep learning classification method have a slower convergence rate of error rate, and the convergence time exceeds 0.7s. Too slow convergence rate on the one hand can easily reduce the efficiency of the classification algorithm. On the other hand, it will also affect the final data classification accuracy; from the error rate convergence curve of the k-means Bayes classification method, it can be observed that the classification error rate of big data converges to zero at 0.29s, and the convergence control The effect is significantly better than the two traditional methods.

Conclusion

Real-time security protection and classified management of the privacy data of cloud storage users is an important content of cloud computing. For the valuable data, redundant data or attacking code and data are often mixed. Therefore, under cloud storage conditions, Based on the Bayes algorithm optimized by k-means, it can provide cloud storage users with more complete encryption management, data backup, and network access control. From the results of the simulation analysis,

it can be seen that the k-means Bayes classification method proposed in the article has better classification performance and ensures the accuracy and efficiency of big data processing in the cloud storage environment.

References

[1] Feng Guilan, Li Zhengnan, Zhou Wengang. Review of the Research on Big Data Analysis Technology in the Network Field [J]. Computer Science, 2019 (6): 1-20

[2] Nie Qingbin, Huo Minxia, Cao Yaoqin, et al. Research on TBCSA-ACO Algorithm in Cloud Computing Task Assignment [J]. Microelectronics & Computer, 2016, 33 (6): 53-58.

[3] Li Ani, Zhang Xiao, Zhang Boyang, et al. Research on performance evaluation methods of public cloud storage systems [J]. Journal of Computer Applications, 2017, 37 (5): 1229-1235.

[4] Wang Chunjuan. Research on Cluster Job Scheduling Algorithm Based on Hadoop Cloud Platform [J]. Bulletin of Science and Technology, 2018 (9): 158-163

[5] Chen Yuxiang, Hao Yao, Zhao Yue, et al. Secure storage exchange technology for manufacturing big data [J]. Application of Electronic Technique, 2019 (12): 38-41

[6] Chen Chen. Research on Library Unstructured Big Data Analysis and Decision System Based on Hadoop [J]. Information Science, 2017 (01): 26-30.