Design of data mining model based on improved manifold learning algorithm in cloud computing environment

ZHAO Zhan-kun

Hebei Software Institute, Baoding Hebei 071000

Keywords: cloud computing; Data; mining model

Abstract. Efficient data mining model design for a large database in the cloud computing environment is studied. For large databases efficiently mining problem, an efficient data mining model in the cloud computing environment based on improved manifold learning algorithms is proposed. The use of nonlinear manifold learning algorithms is able to reduce dimensionality of data vector feature in cloud computing environments, through characteristic extraction module to preprocess data, improved classical manifold learning algorithm is adopted to increase the distance between the data of sample spread intensive area and shorten the distance between the data of sample spread sparse area, prompting even overall distribution of sample database under cloud computing environment, so as to achieve accurate mining for efficient data in cloud computing environment. The experimental results show that the proposed method can accurately mine target data under cloud computing environments, with high efficiency and precision.

Introduction

Cloud computing is an Internet-based supercomputing method, and can be used to perform correlation analysis for data in large databases, to ensure the efficient use of large databases [1]. Therefore, looking for effective data mining methods for large databases, has a very broad space for development in the field of data processing [2]. Due to the amount of data in the large database is gradually increasing, making data in large database presents a strong diversity under cloud computing attributes environment, leading to during the process of data mining in large database based on the traditional tree-based PSO method, the adopted data main element characteristics and associated features have significant volatility, resulting in unable to obtain accurate data mining results [3].

In response to these drawbacks, an efficient data mining model in the cloud computing environment based on improved manifold learning algorithms is proposed. The use of nonlinear manifold learning algorithms is able to reduce dimensionality of data vector feature in cloud computing environments, through characteristic extraction module to preprocess data, improved classical manifold learning algorithm is adopted to increase the distance between the data of sample spread intensive area and shorten the distance between the data of sample spread sparse area, prompting even overall distribution of sample database under cloud computing environment, so as to achieve accurate mining for efficient data in cloud computing environment[4-6]. The experimental results show that the proposed method can accurately mine target data under cloud computing environments, with high efficiency and precision.

The principle of the efficient data mining model in the cloud computing environment based on improved manifold learning algorithms

Problems and optimization methods for manifold learning algorithm.

Manifold learning method aims to find intrinsic regularity of the large database distribution, its main research is points in the large database observation space can joint into a manifold in observation space with the function of a handful of independent variables, if curly manifold in observation space can be effectively expanded or inner main variables can be found, then, the dimension of the database can be reduced. Manifold is the basis of differential geometry, essential is topological space which the area can be coordinated, and can be used as a non-linear universal of

Euclidean space. Manifold learning algorithms considers data structure is linear within a regional context, or the point within a range of area located at same ultra-plane, so that a random point can be described by linear combination of its neighboring points.

Improved manifold learning algorithms adopted for efficient data mining.

The traditional manifold learning algorithms is appropriate for learning in the non- closed linear area, the sampling for samples must be sufficiently smooth, and it is more sensitive to noise, where the number of neighbors k is regarded as an important parameter, its value range is particularly important. Under normal circumstances, manifold learning algorithm assumes the sample extraction library dispersed evenly and continuously, but because of differences of data attribute in a large database, it is possible to be smaller or larger within a certain period of time, and the overall is distributed unevenly. Within the scope of the sample distribution sparsely, the local neighborhood consisting of k nearest neighbors is larger than the local neighborhood consisting of k nearest neighbors within the scope of the sample distribution densely., and the choice of selecting neighbors within ε radius are too rigid. Thus, combining the advantages of the method described above, a new method of selecting the number of k neighbors is proposed.

During large databases efficient data mining process, the number of neighbors constituted by K_{base} and K_{add}, i.e. $K=K_{\text{base}}+K_{\text{add}}$, where K_{base} represents the base value, K_{add} is representative of additional value. First, based on experience to set the value of K_{base}, then calculating the distance between sample points such as X_i and its K_{base}-th neighbor, and multiplied by a constant h. The answer obtained by multiplying is considered as the value of radius ε, so as to acquire the neighbor domain of ε. All K_{base} closed to X_i are data sample points under cloud computing environment, treated as X_i's neighbors, and in the near field of ε and X_i spacing is greater than the distance between sample points X_i and K_{base}-th neighbor points used to determine K_{add}. The following computation contains reconstruction error ε of K_{add} additional neighbors, defined as:

$$\varepsilon(K)=\sum_{I=1}^{N}\left|X_i-\sum_{i=1}^{N}W_{ij}X_j\right|^2 \qquad (1)$$

Since K_{add} for the different data sample point under cloud computing environment varies, therefore ε can be used as a function of K. The optimal number of neighbors can be selected in manifold learning algorithm is to make the reconstruction function to obtain the minimum K. According to results of previous studies, the best value of K_{add} is 6, a constant value is 1.1.

In the condition of selecting the number of optimized neighbors as K, the distance definition in the traditional manifold learning algorithms, is introduced into the proposed algorithm. New distance definition can increase the distance between data samples in sample spread dense regions under the cloud computing environment, shorten the distance between the samples in sample spread sparse region between, prompting the uniform overall distribution of database under cloud computing environments, reducing the interference of K value selection on data mining results under the cloud computing environment. The process of optimized manifold learning algorithms based on the number of new neighbors K is described as follows:

First, for all sample points X_i in high-dimensional space of data under cloud computing environments, K_{base} neighbors are given (usually take $K_{base}=7$), the distance of X_i and K_{base}-th neighbors can be obtained, and multiplied by h (usually take $h=1.1$) to acquired radius parameters of ε neighborhood, the sample points in this area except K_{base} neighbors are used as basis for selecting the value of K_{add}, according to the reconstruction error function equation, K value which can make it smallest is selected as the number of neighbors. All distances are chosen by the distance definition from the traditional manifold learning algorithm, namely:

$$d_{ij} = \frac{\left\| X_i - X_j \right\|}{\sqrt{M(i)M(j)}} \quad (2)$$

Where, $M(i)$ and $M(j)$ represent the mean distance between X_i , X_j and their neighbors K separately.

Secondly, the distance weight W_{ij} from data point X_i in large databases to the neighbor point under cloud computing environment is calculated, that is minimization:

$$\varepsilon(W) = \sum_{i=1}^{N} \left\| X_i - \sum_{j=1}^{N} W_{ij} X_j \right\|^2 \quad (3)$$

In which, assuming j X_j is not a neighbor to i X_i, then $W_{ij} = 0$; $\sum_{j=1}^{N} W_{ij} = 1$

Then, through the weight W_{ij} of high-dimensional space sample X_i of data under cloud computing environments and its neighbor X_j to derive target data in low-dimensional embedding space. Since data in large databases under cloud computing environment can maintain local linear structure of a high dimensional space in the low-dimensional space, and the weights W_{ij} only represent partial information, d therefore weights W_{ij} is fixed to ensure that the loss of the following function is minimalized:

$$\phi(Y) = \sum_{i=1}^{N} \left\| Y_i - \sum_{j=1}^{N} W_{ij} Y_j \right\|^2 = tr(Y^T M Y) \quad (4)$$

Where, $\sum_{i=1}^{N} Y_i = 0$ and $\frac{1}{N} \sum_{i=1}^{N} Y_i Y_i^T = I$, $M = (I - W)^T (I - W)$ is the minimized solution of the equation above, which composed of matrix consisted of corresponding eigenvectors of some maximum eigenvalue of matrix M. Assuming $d+1$ is the eigenvector corresponded by the smallest eigenvalue in M, and the corresponding minimum eigenvector is abandoned, the remaining d eigenvectors makes the matrix, then, the sample value of data in the database of low dimensional space under cloud computing environment can be obtained, eventually achieve a comprehensive data mining under cloud computing environment.

Experimental results analysis

In order to demonstrate the superiority of the proposed data mining method based on improved manifold learning algorithm under cloud computing environments, there is the need for a single experiment. During the experiment, with the traditional algorithm and the proposed algorithm to conduct 8 times data mining experiments in the large database under cloud computing environment, the mining accuracy curve of the two methods obtained as shown in Figure 1. By Figure 1, it can be known that compared to the traditional method, the proposed method used for data mining under cloud computing environment, nearly 5 percent higher for the accuracy, demonstrated that this method has a higher superiority.

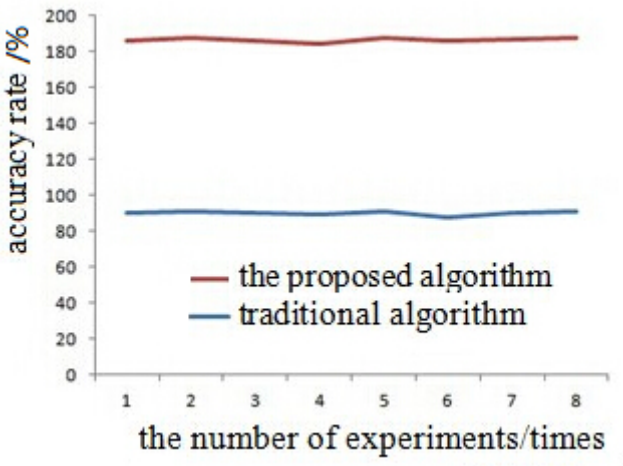

Figure 1 Mining accuracy comparison of two methods

Conclusion

An efficient data mining model in the cloud computing environment based on improved manifold learning algorithms is proposed. The use of nonlinear manifold learning algorithms is able to reduce dimensionality of data vector feature in cloud computing environments, through characteristic extraction module to preprocess data, improved classical manifold learning algorithm is adopted to increase the distance between the data of sample spread intensive area and shorten the distance between the data of sample spread sparse area, prompting even overall distribution of sample database under cloud computing environment, so as to achieve accurate mining for efficient data in cloud computing environment. The experimental results show that the proposed method can accurately mine target data under cloud computing environments, with high efficiency and precision.

References

[1] Sun Jigui, Liu Jie, Zhao Lianyu. Clustering Algorithms Research [J]. Journal of software, 2008, 19(1):48-61

[2] Turney PD.Learning algorithms for keyphrase extraction [J].Information Retrieval, 2000, 2(4):303-336

[3] Meila M,Xu L.Multiway cuts and spectral clustering[R].U.Washington Tech Report. 2003.2-32-35

[4] LIU Min,CHEN De-gang, WU Cheng,et al. Reduction method based on a new fuzzy rough set in fuzzy information system and its applications to scheduling problems[J].Computers & Mathematics with Applications,2005,51(9-10):1571-1583.

[5] Zhuang Like, Kou Zhongbao, Zhang Changshui. Session identification based on time intervals in Web log mining [J]. Journal of Tsinghua University, 2005, 45(1):115-118.

[6] Hou Yali, et al. Data Preparation for Web Log Mining [J]. Journal of Hebei University (natural science edition), 2005, 2(25): 202-206.