Eng Int Syst (2021) 5: 323–330

Engineering
Intelligent Systems

Massive High-Dimensional Big Data Feature Selection Algorithm in a Cloud Computing Environment

Xiaochang Zheng*

Informationand Network Center, Minnan Normal University, Zhangzhou 363000, China

Due to the large number of redundant features in the high-dimensional space, the distance between each sample point is almost equal in the whole feature space. Current algorithms cannot retain and utilize the key features of the original data to their maximum effect in order to get a higher classification prediction result. A high-dimensional and big data feature selection algorithm based on PageRank is proposed. According to the set of information entropy threshold, a large number of high-dimension and big data are screened in the cloud computing environment, the attributes of the almost useless raw data information are discarded, and the dimensionality reduction processing is completed. On the basis of dimensionality reduction processing, the sample is determined. Each data feature in the sample is regarded as a network node, and the edge of the node is created according to the mutual information. The PageRank algorithm is used to evaluate the global redundancy of the network nodes, and the nodes are sorted according to the evaluation criteria, where the first g features of the sequence are the optimal feature subset. The experimental results show that the proposed method can not only get a good visual division result, but can also reach a higher accuracy of classification prediction.

Keywords: Cloud Computing Environment; Massive; High-Dimensional; Big Data; Feature Selection.

1. INTRODUCTION

For a long time, people have wanted to reveal the objective laws of things and phenomena hidden in complex imagery. In order to provide more information and more complete information, people constantly develop new observation tools, new observation techniques, and produce more and more large-scale high-dimensional data. This data is not only large in number, but also has many separate features. Knowledge refers to a form of data that has been processed and refined. It plays a vital role in human understanding and the transformation of the objective world. The traditional method of knowledge acquisition is to build a cognitive model directly or intuitively from the data in order to obtain useful information that people can understand. This manual acquisition method is very useful and efficient for a small amount of data, but when facing a large amount of datathe limitations of this method become clear - that it cannot meet the needs of the rapid development of information. This is the phenomenon of "data explosion and poor knowledge". Therefore, finding ways to effectively process data and determine useful knowledge from massive data is a difficult problem facing people. This is also the main focus of the research content concerning information processing. Feature selection and feature extraction of high-dimensional data are a key link in pattern recognition. From the middle of last century, many researchers began to study this aspect, put forward various theories and methods, and accumulate a large number of research results. Today, pattern recognition is widely used in biometrics, biological information classification and other tasks, which show a broad development prospect (Lin et al., 2016).

*Corresponding Author e-mail: waw5217@126.com

Many superior data classification learning algorithms have been put forward at present, they can be well qualified for small scale data and can quickly discover the regularity of data; however, with the emergence of new technology, the real-world data set is accumulated and developed in a large direction, which requires more resources and new more efficient methods. Digging or learning methods must also adapt to this situation, in order to efficiently handle large-scale data and eventually acquire useful knowledge. Feature selection is a hot research topic in the fields of statistics, data mining and machine learning. At first, the concept was put forward to solve the problem of large-scale data computation. However, due to the limitation of technology at that time, the dimensionality of processing was insufficient. With the continuous emergence of new technology, the actual application of data began in a more large-scale direction, and this brought a severe test of the traditional methods of feature selection (Zhang et al., 2016).

Wang et al. proposes a massive high-dimensional big data feature selection algorithm based on random matrix theory (RMT) (Wang et al., 2017). In the cloud computing environment, the singular values of the massivehigh-dimension and big data correlation matrix which conform to the random matrix prediction are removed, and the correlation matrix and the number of selected features is obtained after the de-noising process. Then on the correlation matrix was decomposed by the singular value and the correlation between the features and the classes is obtained by the decomposition matrix. Feature selection is then completed by redundancy. Ji et al. proposes a massivehigh-dimensional data feature selection algorithm based on granular fusion framework (Ji et al., 2016). Using the idea of BLB (Bag of Little Bootstrap), the original mass data sets are granulated into a small scale data subset (particle). A multiple set of self-help subsets is constructed on each particle to achieve the particle feature selection and each particle feature selection node is implemented. According to the weight and ranking, the result of the ordered feature selection of the original dataset is obtained. Wang et al. proposes a massive high-dimensional big data feature selection algorithm based on Granular Computing and discriminating ability (Wang et al., 2017). By using the stratified sampling technique in statistics, the massivehigh-dimensional big data sets in the cloud computing environment are split into multiple sample subsets, the ability to distinguish the data attributes on each particle is calculated, the theory of Niche Immune Optimization is fused, and the classification approximation standard of the attribute set is introduced to reduce the affinity of the immune optimization as the data attribute. The Niche Immune sharing mechanism is then generated, and the massive high-dimension and big data attributesare reduced. Finally, the massivehigh-dimensional and big data feature selection model is established in the cloud computing environment, and the feature selection is realized. Zhou et al. proposes a massive high-dimensional big data feature selection algorithm based on data driven k-nearest neighbor mutual information (Zhou and Qiao, 2017). The algorithm extends the data driven k-nearest neighbor method to the estimation of mutual information between large and big data feature variables in the cloud computing environment. The optimal sorting of all features is given by the forward accumulation strategy, and the unrelated features are eliminated according to the number of presupposition independent features, and then the backward crossover strategy is used to identify and eliminate redundant features. Finally, the optimal subset of strong correlation features is obtained and the massive high-dimensional big data feature selection is completed (Ahmed, 2019; Jugal and Gupta, 2019; Tan et al., 2019).

The above algorithm cannot effectively reduce or eliminate the influence of redundant features and unrelated features, so as to retain and utilize the key features of the original data to the maximum extent, and this seriously affects the accuracy of classification prediction. A PageRank based algorithm for massive high-dimensional big data feature selection is proposed, and the specific research steps are given.

The first step is to reduce the dimension of massive high-dimensional big data in the cloud computing environment.

The second step is to determine the size of the massive data samples in the cloud computing environment.

The third step is to use the PageRank algorithm to evaluate the global redundancy of network nodes, sort the nodes according to the evaluation criteria and determine the first g features of the sequence, that is, the optimal subset.

The fourth step is to prove the proposed method can obtain a higher classification accuracy when compared with other feature selection algorithms.

2. RESEARCH ON MASSIVE HIGH-DIMENSIONAL BIG DATA FEATURE SELECTION ALGORITHM IN CLOUD COMPUTING ENVIRONMENT

2.1 Dimension Reduction of Massive High-Dimensional Big Data Feature in Cloud Computing Environment

The data features of high-dimension and big data are immense. When PCA is used to reduce the dimension of the data, the memory consumption is large and it is very time consuming. The available memory of a standard PC is insufficient to perform the memory analysis and calculation in a timely manner. In this application, in order to obtain the ideal results, we need to improve the PCA, compress the data as much as possible, and reduce the dimension of the data, while keeping the memory occupancy and running time within the application requirements.

According to the meaning and application of information entropy, information entropy can be used to reduce the dimension and explain the following: The greater the information entropy of the feature, the greater the amount of information it contains, and more of the information should be retained; If the information entropy of a feature is lower, less of the data is contained in the feature, and the feature is made into a feature extraction. When taking or reducing dimensions, the feature is not included in the option. According to the entropy of information, the feature is either to be excluded or

to be retained, which can be determined by the following two schemes (Kang et al., 2016):

(1) The information entropy is set to select the features. According to the importance of the feature to the application analysis, the entropy values of the useful and the useless feature information has an obvious dividing point. In this case, the entropy value is done at the boundary point.

(2) To preserve the features of a given proportion. Expressions are used to calculate the proportion of selected features in the original data.

$$\frac{\sum_{i=1}^{k} H(i)}{\sum_{j=1}^{k} H(j)} \geq threshold \tag{1}$$

Among them, $H(i)$ and $H(j)$ respectively denote the information entropy of massive high-dimensional big data sets with i and j attributes in the cloud computing environment; *threshold* indicates thresholds; k is constant.

Information entropy can be used to reduce the dimension of massive high-dimensional big data in the cloud computing environment, so as to reduce the time and space complexity of the subsequent processing algorithm.

The improved PCA algorithm will set the information entropy threshold δ to filter the big data in the cloud computing environment, and discard the features of the almost useless original data information (Li et al., 2017). The specific operation process is described as follows:

(1) The massive high-dimensional big data in the cloud computing environment is transformed into a matrix $U_{n\times m}$, where m represents the number of samples and n represents the number of features.

(2) By using Equation (2) to calculate the information entropy $H(a_i)$ of each attribute a_i in the samples:

$$H = E\left[-\log p_i\right] = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

Among them, p_i represents the probability of massive high-dimensional big data sample attributes to a_i features in cloud computing environment; $E\left[-\log p_i\right]$ represents the statistical mean of $-\log p_i$.

According to the above calculation, the feature selection is compared with the information entropy threshold δ set by the improved PCA algorithm. If the following conditions are satisfied, the data of the attribute is a_i, and all the attributes of the a_i are put into the set; otherwise, all the attributes of the a_i can not be put into the set.

$$H(a_i) > \delta \tag{3}$$

(3) The centralized matrix $X_{n\times m}$ of massive high-dimensional big data samples in the cloud computing environment can be obtained. The formula is as follows:

$$X_{n\times m} = A - repmat(mean(A, 2), 1, m) \tag{4}$$

In this, $repmat()$ represents a function in MATLAB.

(4) Calculate the mean m_i and centralization mean x_{ij} of massive high-dimensional big data samples in the cloud computing environment:

$$m_i = \frac{\sum_{j=1}^{m} u_{ij}}{m} \tag{5}$$

$$x_{ij} = u_{ij} - m_i \tag{6}$$

In the formula, u_{ij} represents the value of attribute i in massive high-dimensional big data sample j in the cloud computing environment.

(5) According to the above calculation, the covariance matrix Cov of different attributes can be obtained. The formula is as follows:

$$Cov = \begin{pmatrix} \mathrm{cov}(x_{1i}, x_{1i}) & \mathrm{cov}(x_{1i}, x_{2i}) & \cdots & \mathrm{cov}(x_{1i}, x_{mi}) \\ \mathrm{cov}(x_{2i}, x_{1i}) & \mathrm{cov}(x_{2i}, x_{2i}) & \cdots & \mathrm{cov}(x_{2i}, x_{ni}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cov}(x_{ni}, x_{1i}) & \mathrm{cov}(x_{ni}, x_{2i}) & \cdots & \mathrm{cov}(x_{ni}, x_{mi}) \end{pmatrix} \tag{7}$$

The eigenvector *eigenVector* and eigenvalue *eigenValue* of Cov can be obtained from Equation (7).

(6) Select the transform base by selecting the k eigenvectors corresponding to the largest k eigenvalues as the column vectors to form the feature vector matrix $V_{n\times k}$.

(7) In the available cloud computing environment, the dimensionality reduction results of massive high-dimensional big data are as follows:

$$S = V_{n\times k} \times X_{n\times m} \times Cov \tag{8}$$

2.2 Determination of Sample Size for Massive High-Dimensional Big Data Sets in the Cloud Computing Environment

After reducing the dimension of the large dimension and big data in the cloud computing environment, a common method to determine the size of the sample is given. By redefining some of the variables in the calculation formula, it is extended to a size that can determine the size of the data centralized sampling (Li et al., 2016).

Suppose S represents a data table after a reduction of the high-dimension and big data in the cloud computing environment; N represents the size of the table, and the calculation formula of the size M' of the sample of the table S is as follows:

$$M' = \frac{Z^2 \times \sigma^2}{E^2} \tag{9}$$

In the formula, σ represents the standard deviation on the data table S; Z represents a quantile value under the confidence interval; and E represents the marginal error.

According to the above calculation, it can be seen that the size of the massive high-dimensional big data sample set in the cloud computing environment is obviously not related to the size of the estimated data table, but a large number of research examples show that when the size of the sample size

M' exceeds the estimated size of 5% of the estimated data set, it is necessary to make a further correction to the M', through the introduction of the N. The correction coefficient of the finite population is further revised for the sample size M'.

$$M = \frac{M' \times FPC \times N}{M' + N + 1} \tag{10}$$

Introducing the dissimilarity coefficient μ instead of standard deviation σ, and assuming that U represents the domain of the data table S, then:

$$\mu = \frac{\sum_{i=1}^{|U|} \sum_{j=1}^{|U|} c(x_i, y_i)}{|U|^2} \tag{11}$$

$$c(x_i, y_i) = \begin{cases} 1, & x_i \neq y_i \\ 0, & x_i = y_i \end{cases} \tag{12}$$

In the formula, $x_i, y_i \in U$ represents any two data in the domain U; and i and j represent two different attributes of the data.

The $x_i \in U$ in Equation (12) represents a one-dimensional vector, but in fact data in the data table S is usually multi-dimensional, so $c(x_i, x_j)$ is extended to a multi-dimensional vector and is recorded as$c_d(x_i, x_j)$. In which, d represents the dimension; set $S = (U, C \cup D)$, C is used to describe the set of sample attributes; D represents class labels; $a \in C$, $x_i = (a_1(x_i), a_2(x_i), \cdots, a_{|C|}(x_i))$, then the formula for the expanded multi-dimensional vector $c_d(x_i, y_i)$ is as follows:

$$c_d(x_i, x_j) = \sum_{\kappa=1}^{|C|} \xi(a_\kappa(x_i), a_\kappa(x_j)) \tag{13}$$

Among them, κ is a constant, and the formula for function $\xi(a_\kappa(x_i), a_\kappa(x_j))$ is:

$$\xi(a_\kappa(x_i), a_\kappa(x_j)) = \begin{cases} 1, & a_\kappa(x_i) \neq a_\kappa(x_j) \\ 0, & a_\kappa(x_i) = a_\kappa(x_j) \end{cases} \tag{14}$$

Based on the above analysis, the dissimilarity coefficient μ in table S can be reformulated as follows:

$$\hat{\mu} = \frac{\sum_{i=1}^{|U|} \sum_{j=1}^{|U|} c_e(x_i, x_j)}{|U|^2} \tag{15}$$

Correspondingly, the size of massive high-dimensional big data samples in the cloud computing environment can be represented as:

$$\hat{M} = \frac{Z^2 \times \hat{\mu}}{E^2} \tag{16}$$

2.3 Feature Selection Algorithm Based on PageRank

On the basis of determining the size of massive high-dimensional and big data samples under the cloud computing environment, each data feature is regarded as a network node, and the edge of the node is created according to the mutual information. The PageRank algorithm is used to evaluate the global redundancy of the network nodes, and the nodes are sorted according to the evaluation criteria. The first g features of the sequence are optimal as the feature subset.

(1) Computation of redundancy between features

In order to solve the shortcoming of categorical features without taking into account the redundancy of measurement, a feature redundancy metric method is proposed, which takes into account the redundancy between the features of a given data mining task (Sun et al., 2016). A correlation measure is defined, which is used to quantify information redundancy between feature L_i and feature L_s under the condition of class feature F.

$$R(L_i; L_s) = \frac{I(F; L_i) + I(F; L_s) - I(F; L_i, L_s)}{H(F)} \tag{17}$$

$$I(F; L, L_s) = I(F; L_s) + I(C; L_i | L_s) \tag{18}$$

Among them, $I(F; L_i)$ represents the common information between the category feature F and the feature L_i; $I(F; L_s)$ represents the common information between the category feature F and the feature L_s; $I(F; L_i, L_s)$ expresses the common information between the category feature F and the feature L_i and L_s; and $H(F)$ expresses the information entropy of the category feature F.

The above measures show excellent performance in many experiments, but in a few cases this measure may have a shortcoming, that is, when two of the three of F, L_i, and L_s are independent, $R(L_i; L_s) \leq 0$

According to Equations (17) and (18), it can be obtained:

$$R(L_i; L_s) = \frac{I(F; L_i) I(C; L_i | L_s)}{H(F)} \tag{19}$$

It is known that:

$$I(F; L_i; L_s) = I(F; L_i) I(C; L_i | L_s) \tag{20}$$

Then:

$$R(L_i; L_s) = \frac{I(F; L_i; L_s)}{H(F)} = \frac{I(F; L_s) I(F; L_i)}{H(L_s) H(F)} \tag{21}$$

Among them, $H(L_s)$ represents the information entropy of the feature L_s.

The above measurement features redundancy between class L_i and L_s. If $R(L_i; L_s) = 0$, the feature L_i and L_s are independent of category feature F, and, on the contrary, the greater the value of the $R(L_i; L_s)$, the greater the redundancy of the feature L_i and L_s on the category feature F(Komeili et al., 2018)

The following transformation is made to Equation (21):

$$R(L_i; L_s) = \frac{\frac{I(F; L_s)}{H(L_s)} I(L_i; L_s)}{H(F)} \tag{22}$$

Among them, $I(L_i; L_s)$ represents the common amount of information between L_i and L_s

From Equation (22), we can see that the redundant $R(L_i; L_s)$ between the feature L_i and L_s on the category feature F is estimated by the feature L_s. In the same way, $I(L_s; L_i)$ is estimated by the feature L_i, usually $R(L_i; L_s) \neq R(L_s; L_i)$, and sometimes the values of the two are very different.

(2) PageRank evaluation feature global redundancy

Take each feature as a node with a weighted network, and note that Q represents the number of features in the sample set, then the network is composed of Q nodes, and the i node q_i corresponds to the i feature L_i, and in the directed weighted network, for each node to q_i and q_j, if $R(L_i; L_j) \neq 0$, there is a side from node q_j pointing node and the weight of the edge is $R(L_i; L_j)$. Similarly, if $R(L_j; L_i) \neq 0$, then there is an edge from node q_i to node q_j, and the weight of the edge is $R(L_j; L_i)$. In a directed additive network, a node q_i will transfer its own importance to all the nodes it points to, and this proportion is specific to $R(L_s; L_i)/\sum_{L_j \in out_i} R(L_j; L_i)$, in which out_i represents the set of the corresponding features of all nodes pointing to the node q_i in the network (Zhang et al., 2016).

There are two possible problems in directed weighted networks corresponding to actual datasets. One of these is that one node may not point to any other node; the other possibility considers two nodes, where there are two nodes pointing to each other, but neither of the two points to any other node, and one or more other nodes point to one of them (Lee et al., 2016). The first problem is called the level leak, and the second problem is called the level sunk. These two problems can be solved by the processing of the PageRank. The feature redundancy evaluation formula can be expressed as follows:

$$PR(L_s) = \frac{1-k}{Q} + k \sum_{i=1, i \neq s}^{Q} PR(L_i) \times \frac{R(L_s; L_i)}{\sum_{L_j \in out_i} R(L_j; L_i)} \tag{23}$$

The upper formula is expressed in vector and matrix form as:

$$Rank = \left[kH + (1-k) \times \left[\frac{1}{Q} \right]_{Q \times Q} \right] Rank \tag{24}$$

The $Rank$ in the upper formula is a $Q \times 1$ vector, the i feature of the vector is $PR(L_i)$; H represents a random square of the directed graph corresponding to the network, and the calculation formula is as follows:

The entire vector $Rank$ in the sample set satisfies $\|Rank\| = 1$. The $PageRank$ value of all features can be obtained by an iterative calculation of Equation (24) until the termination condition is satisfied. In the process of creating a directed network, it is based on the redundancy of the feature L_i and the feature L_s on the redundancy degree $R(L_i; L_s)$ and the edge weight is $R(L_i; L_s)$, so the larger the $PageRank$ value of a special feature is, the greater the redundancy. (García-Torres et al., 2016).

(3) Feature score and selection of optimal specific subset

According to the above PageRank algorithm, the correlation between L_i and F is estimated by using the mutual information $I(L_i, F)$ between the feature L_i and the category feature F, and $I(L_i, F)$ is normalized by the following methods:

Let $I_{fc} = \left[I(L_1, F), I(L_2, F), \cdots, I(L_Q, F) \right]$ and set up a mapping *func*:

$$I(L_i, F) \to func(I(L_i, F)) = I(L_i, F) / \sum_{l=1}^{Q} I(L_l, F) \tag{26}$$

Among them, $l = 1, 2, \cdots, Q$

In the cloud computing environment, the correlation formula of each feature in the massive high-dimensional big data sample set is as follows:

$$I_{norm} = \left[\frac{I(L_1, F)}{\sum_{l=1}^{Q} I(L_l, F)}, \frac{I(L_2, F)}{\sum_{l=1}^{Q} I(L_l, F)}, \cdots, \frac{I(L_Q, F)}{\sum_{l=1}^{Q} I(L_l, F)} \right] \tag{27}$$

The feature redundancy evaluation results $PR(L_i)$ and I_{norm}, which are obtained by the above calculation, are used to give the following feature evaluation criteria J_{fscn}. According to the evaluation criteria, all the features of the sample set are arranged from large to small in accordance with the evaluation values, and the first g feature of the sequence is the optimal subset of the features (Yu et al., 2017). The calculation formula is described as follows:

$$J_{fscn} = I_{norm} - PR(L_i) \tag{28}$$

3. CONSTRUCTION OF EXPERIMENTAL ENVIRONMENT AND ANALYSIS OF RESULTS

In order to test the effectiveness of the PageRank based large dimension and big data feature selection algorithm, the experiment compares the proposed algorithm with the algorithms proposed in literature. These four feature selection algorithms are independent of the classifier. In order to get the accuracy of classification prediction for various feature selection algorithms, the nearest neighbor classifier is selected for prediction. The test results of the four feature selection algorithms on UCI dataset are given respectively. All experiments are run on a notebook with a 2.26 GHZ CPU and are simulated in MATLAB software.

3.1 Comparison of UCI Data Sets

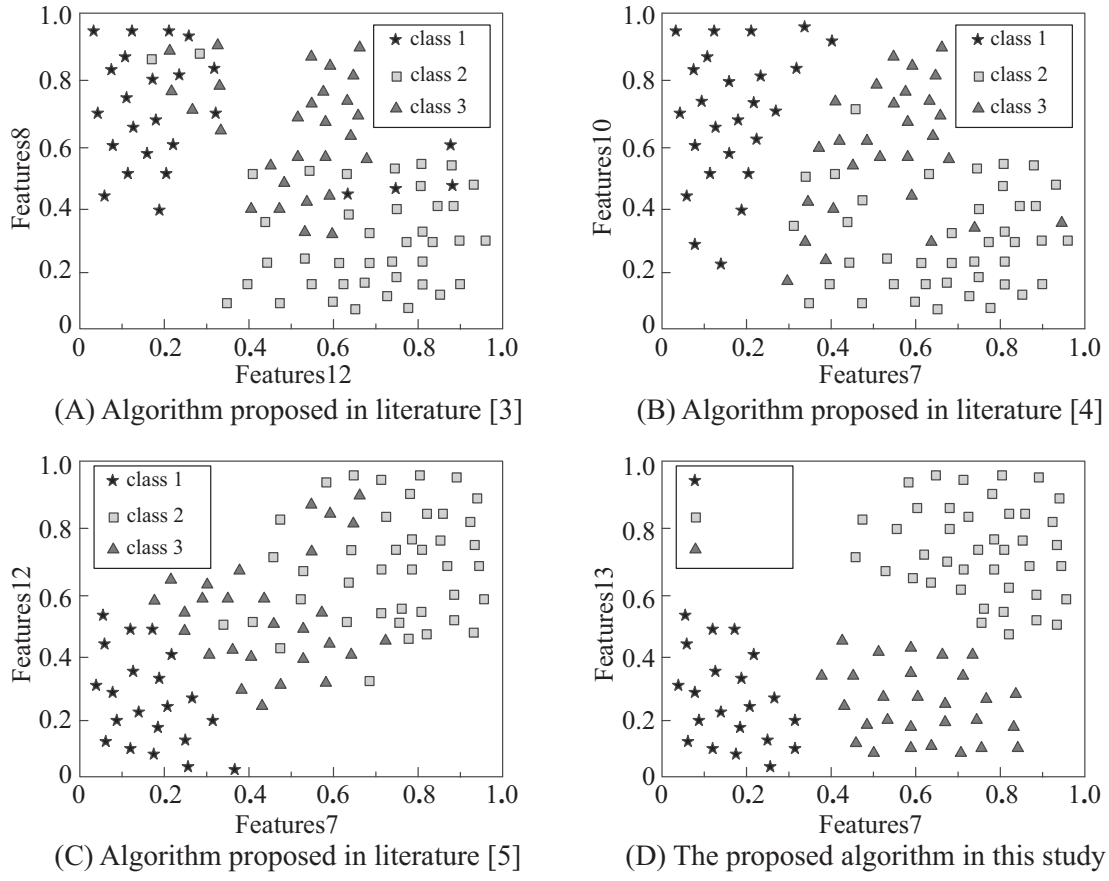
4 feature selection algorithms are compared using the UCI dataset. From the UCI dataset, 6 sets of data are selected for test comparison between the algorithms, as shown in Table 1.

According to the experimental results of the 4 feature selection algorithms, the importance ranking of the 6 data sets is obtained. Figure 1 and Figure 2 show a two-dimensional visualization of the Wine and Breast-diagnostic datasets, respectively. The X-axis co-ordinates and the Y-axis co-ordinates represent the two most important features of each data set. It can be clearly seen from the diagram that, for

$$H = \begin{cases} 0 & \frac{R(L_1;L_2)}{\sum_{L_j \in out_2} R(L_j;L_2)} & \frac{R(L_1;L_3)}{\sum_{L_j \in out_3} R(L_j;L_3)} & \cdots & \frac{R(L_1;L_Q)}{\sum_{L_j \in out_Q} R(L_j;L_Q)} \\ \frac{R(L_2;L_1)}{\sum_{L_j \in out_1} R(L_j;L_1)} & 0 & \frac{R(L_2;L_3)}{\sum_{L_j \in out_3} R(L_j;L_3)} & \cdots & \frac{R(L_1;L_Q)}{\sum_{L_j \in out_Q} R(L_j;L_Q)} \\ \frac{R(L_3;L_1)}{\sum_{L_j \in out_1} R(L_j;L_1)} & \frac{R(L_3;L_2)}{\sum_{L_j \in out_2} R(L_j;L_2)} & 0 & \cdots & \frac{R(L_3;L_Q)}{\sum_{L_j \in out_Q} R(L_j;L_Q)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{R(L_Q;L_1)}{\sum_{L_j \in out_1} R(L_j;L_1)} & \frac{R(L_Q;L_2)}{\sum_{L_j \in out_2} R(L_j;L_2)} & \frac{R(L_Q;L_3)}{\sum_{L_j \in out_2} R(L_j;L_3)} & \cdots & 0 \end{cases} \quad (25)$$

Table 1 UCI Data Set Used in the Experiment.

Data set information	Data set size	Dimension	The number of clusters
Iris	180	5	3
Wine	200	12	3
Ionosphere	350	32	2
Heart	270	14	2
Connectionist bench	210	61	2
Breast-diagnostic	560	33	2



(A) Algorithm proposed in literature [3]
(B) Algorithm proposed in literature [4]
(C) Algorithm proposed in literature [5]
(D) The proposed algorithm in this study

Figure 1 2D Visualization of Wine Data Sets Obtained From Four Different Algorithms.

both the Wine data sets of three data clusters and the Breast-diagnostic data sets of two data clusters, the massive high dimensional big data feature selection algorithms based on the PageRank receive good visualization results. In subsequent experiments, the accuracy of classification prediction will be further evaluated.

On the basis of the above experiments, 6 UCI data sets are randomly divided into ten parts, and the accuracy rate of the classification prediction with the nearest neighbor classifier is tested by the cross-validation method. In order to ensure the fairness of the experimental comparison, 20 repeated experiments were carried out for all the algorithms, and the average values of the test results were compared. Figure 3 shows the transformation of classification accuracy with the increase in the number of data sets. From Figure 3, we can see that the proposed algorithm achieves the highest prediction accuracy in most cases. Table 2 gives a detailed description of the highest classification accuracy (%) and the number of

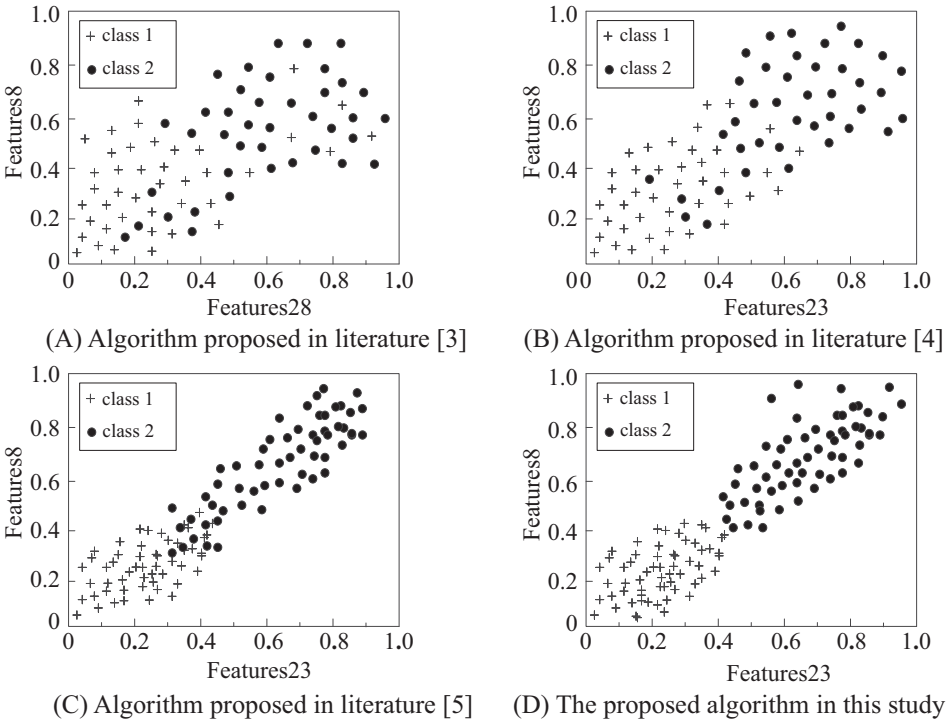


Figure 2 2D Visualization of Breast-Diagnostic Data Sets Obtained From Four Different Algorithms.

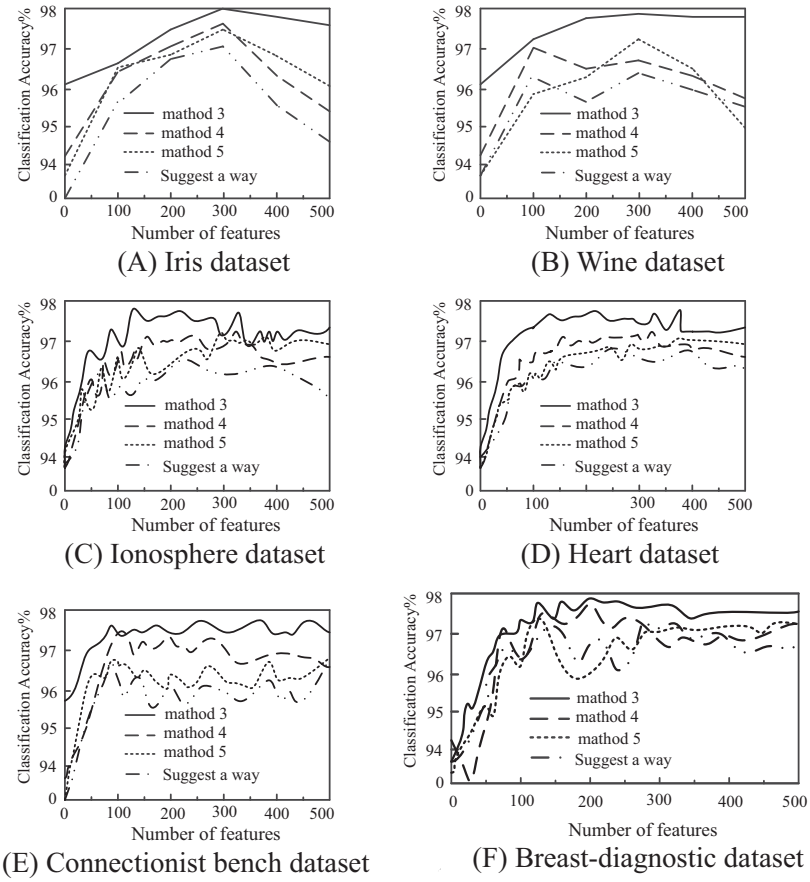


Figure 3 Feature Selection Algorithm Increases the Accuracy of Classification Prediction With the Number of Data Sets.

corresponding features (in parentheses) of the different feature selection algorithms on the UCI dataset. The experimental results show that the PageRank based high-dimensional big data feature selection algorithm has better classification results than the other three algorithms. At the same time, it also shows that the feature selection algorithm is used to get some

Table 2 The Highest Classification Accuracy and Corresponding Number of Features Obtained by Various Feature Selection Algorithms on UCI Dataset.

	Algorithm in Literature [3]	Algorithm in Literature [4]	Algorithm in Literature [5]	The proposed algorithm
Iris	93.7(3)	94.2(3)	95.6(3)	96.8(3)
Wine	96.7(10)	97.2(10)	97.8(10)	98.1(8)
Ionosphere	89.4(16)	88.5(19)	87.4(22)	95.2(11)
Heart	76.8(11)	78.3(12)	75.4(14)	89.7(9)
Connectionist bench	85.9(58)	86.2(64)	86.9(61)	88.7(43)
Breast-diagnostic	85.2(10)	85.6(16)	86.2(28)	88.3(12)

reasonable subsets of the feature, and a higher classification accuracy can be obtained than by using all of the data features.

4. CONCLUSIONS

In this paper, a massivehigh-dimensional big data feature selection algorithm based on PageRank is proposed. In the MATLAB software environment, the UCI data set is used as the test data set, and three other feature selection algorithms are compared in detail. The results show that the massivehigh-dimensional big data feature selection algorithm based on PageRank has a higher classification result compared with three other kinds of feature selection algorithms, and this can be widely applied to the research of classification, clustering and visualization.

In subsequent research work, we will continue to focus on several extensible aspects of the proposed algorithm. First, it is necessary to point out that the in-depth analysis of the mathematical theory of the algorithm is still insufficient, and the selection interval of the ideal parameters is the key point of the next step. Secondly, in future research work, the class information of the data samples can be introduced into the process of feature selection, and better feature selectionsare able to be selected. Thirdly, the research will consider the correlation between data features, not only based on the fractional equation of a single data feature, but also the correlation information of the feature subset. Finally, we hope to further analyze and verify the feature selection algorithm based on PageRank and further analyze the feature from a mathematical point of view.

ACKNOWLEDGEMENTS

The work was funded by Project of Fujian Provincial Science and Technology Department (Grant No: 2017J01405); Fujian Provincial Department of Education Project of Science and Technology (Grant No: JAT170366); Fujian Provincial Department of Education Project of Science and Technology (Grant No: JAS170258).

REFERENCES

1. Ahmed A. (2019). Effect of Corona on the Wave Propagation along Overhead Transmission Lines. Acta Electronica Malaysia, 3(1): 06–09.
2. García-Torres, M., Gómez-Vela, F., Melián-Batista, B., et al. High-dimensional feature selection via feature grouping: A variable neighborhood search approach. Information Sciences, 2016, 102–118.
3. Ji, S.Q., Shi, H.B., Lv, Y.L., et al. Feature selection algorithm based on granulation-fusion for massive high-dimension data. Pattern Recognition and Artificial Intelligence, 2016, 7: 590–597.
4. Jugal K. G., S.K. Gupta (2019). A Comparative Study of Crowd Counting and Profiling Through Visual and Non-Visual Sensors. Acta Informatica Malaysia, 3(1):04–06.
5. Kang, M., Islam, M.R., Kim, J., et al. A hybrid feature selection scheme for reducing diagnostic performance deterioration caused by outliers in data-driven diagnostics. IEEE Transactions on Industrial Electronics, 2016, 5: 3299–3310.
6. Komeili, M., Louis, W., Armanfard, N., et al. Feature selection for non-stationary data: Application to human recognition using medical biometrics. IEEE Transactions on Cybernetics, 2018, 5: 1446–1459.
7. Lee, J.H., Oh, S.Y. Feature selection based on geometric distance for high-dimensional data. Electronics Letters, 2016, 6: 473–475.
8. Li, F., Zhang, Z., Jin, C. Feature selection with partition differentiation entropy for large-scale data sets. Information Sciences, 2016, 690–700.
9. Li, J., Liu, H. Challenges of feature selection for large data analytics. IEEE Intelligent Systems, 2017, 2: 9–15.
10. Lin, C.K., Zhang, K.Y., Huang, Y.H., et al. Feature selection based on an improved cat swarm optimization algorithm for large data classification. Journal of Supercomputing, 2016, 8: 3210–3221.
11. Sun, S., Peng, Q., Zhang, X. Global feature selection from microarray data using Lagrange multipliers. Knowledge-Based Systems, 2016, 267–274.
12. Tan M., Yuan J.S., Wei M., Zhang Y.Q. (2019). Commercial Complex Intelligence and Program Research. Information Management and Computer Science, 2(1): 04–06
13. Wang, S.P. The simulation research on the accurate extraction of the feature data in mobile internet. Computer Simulation, 2017, 2: 322–325.
14. Wang, Y., Yang, J., Sun, L.F., et al. Feature selection method of high-dimensional data based on random matrix theory. Journal of Computer Applications, 2017, 12: 3467–3471.
15. Yu, Z., Zhang, Y., You, J., et al. Adaptive semi-supervised classifier ensemble for high dimensional data classification. IEEE Transactions on Cybernetics, 2017, 99: 1–14.
16. Zhang, X., Mei, C., Chen, D., et al. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. Pattern Recognition, 2016, 1: 1–15.
17. Zhang, X., Mei, C., Chen, D., et al. Feature selection in mixed data. Pattern Recognition, 2016, 1–15.
18. Zhou, H.B., Qiao, J.F. Feature selection method based on high dimensional k-nearest neighbors mutual information. CAAI Transactions on Intelligent Systems, 2017, 5: 595–600.