Distributed dimensionality reduction of industrial data based on clustering

Yongyan Zhang, Guo Xie, Wenqing Wang, Xiaofan Wang, Fucai Qian, Xulong Du,Jinhua Du
Xi'an University of Technology, Shaanxi Key Laboratory of Complex System Control and Intelligent Information Processing, China
guoxie@xaut.edu.cn

***Abstract*—Large amounts of data are produced in system operation, and how to extract effective information from these data has become an important research topic in the industrial application. Dimensionality reduction is a way to refine the data. Because of the low efficiency of the existing methods, these methods can't discover the internal structure of the data. Regarding these problems, a distributed method of dimensionality reduction based on clustering is proposed, which includes the following steps:(1) Clustering the data into some small classes according to the similarity between the data variables; (2) reducing the dimension of data in a small class after being clustered respectively; (3) merging the data after being reduced dimension; (4) classifying the data after being merged by support vector machine (SVM). The data in the simulation is the test data, and the results show that the methods proposed in this paper are better than the existing dimensionality reduction methods.**

Keywords—data dimensionality reduction; SVM; clustering; classification

I. INTRODUCTION

With the large amount of industrial data generated, and how to extract effective information from exponential growth data has become more and more important. If it analysis the high dimensional data directly, it will be calculated in a large consumption and it is easy to ignore the internal structure of the data. If it can make a dimension reduction of the data before analyzing the data, it will reduce the amount of computation for data processing and it is easy to find the internal structure of the data. Dimensionality reduction is an important tool to extract effective information, and the major methods of dimensionality reduction include: traditional PCA [1], LDA [2], LLE [3], KPCA [4], multi-layer automatic encoder [5], MDS [6-7] and so on. Dimensionality reduction is an important step in data preprocessing, and it is always used in classification, regression [8, clustering [9] and other areas, for example data mining and pattern recognition.

In order to extract feature information from a large number of data effectively, reference [10] proposed a multi-layer encoder which the high-dimensional data mapping to low-dimensional space, and extract the important information effectively to complete the image reconstruction; reference [11] proposed a method of dimensionality reduction SNE, and this method used probability to measure the similarity of data points between high-dimensional and low-dimensional space, assuming that the probability distributions of the low-dimensional space and the high-dimensional are similar, and which used a KL distance to determine the similarity between the high-dimensional space and the low-dimensional space; reference [12] proposed an algorithm named as tSNE which is an improved algorithm for the "crowding" and optimization problems in the SNE algorithm, and replace the conditional probability with joint probability, improving the visualization effect of the data. However, these methods have some limitations, and it is only for a certain type of data. For example, the SNE and tSNE algorithms are only better for visualization in 2D or 3D space, and the calculation of this iterative algorithm is very large; PCA is performed well for the linear relationship, and can't extract information for non-linear data; KPCA is good for non-linear data, but the calculation is so large, and so on.

In order to improve the efficiency of the existing dimensionality reduction methods, the paper proposed a distributed method of combine clustering and dimensionality reduction algorithm. The combination of inherited clustering and dimension reduction are mentioned in the literature[13], and the purpose of dimension reduction is clustering. Now, the clustering method is placed before the dimensionality reduction, and clustering can make the related information into one category, of course it also includes redundant information. The clustering can group related information into one category, of course it also includes redundant information. The algorithm clusters the data variables according to the similarity firstly, and the data is divided into k small class; then reduce the dimension of k small class data after being clustered respectively; and then merge the data after being reduced dimension; finally classify the merged data by SVM. Reference [14] had use the accuracy of classification as the standard for evaluating the effect of dimensionality reduction, and the different dimensionality reduction methods [15] have a great influence on the accuracy of classification. Finding valid combinations from different clustering and dimensionality reduction algorithms, and make the effect is remarkable. The results of simulation show that the methods proposed in this paper are more effective compared with the existing methods of dimensionality reduction.

The rest of this paper is organized as follows. In Section 2, we raised the question, and introduced the detailed process of the algorithm and evaluation method SVM of dimensionality reduction. The numerical simulation and result analysis are

978-1-5386-3758-6/18/$31.00 ©2018 IEEE

described in detail in Section 3. Finally, Section 4 gave the conclusions.

II. DATA DIMENSION REDUCTION BASED ON CLUSTERING

A. Questions raised

When the dimension of the data is high, the problems of data processing become very complicated, and this is so called "dimensionality disaster." The correlation dimension of the data is increases, and the number of related data will be increases exponentially. If the high dimensional data can be expressed in low-dimensional space, the amount of computation will be reduced greatly.

Now, it is exist m column data $\{X_1, X_2, \cdots, X_m\}$, and $X_i = [\mathrm{x}_{1i}, \mathrm{x}_{2i}, \cdots, \mathrm{x}_{ni}]^{\mathrm{T}}$ $i = 1, 2, \cdots, m$. Which m indicates the number of data variables, and n indicates the number of samples. In order to find the structure of hidden in low-dimensional from high-dimensional space, and it is reduce the dimension of m column data, extracting the effective information. Taking the classic PCA algorithm as an example, now it will introduce the process of data dimensionality reduction directly.

Input: sample sets $\{X_1, X_2, \cdots, X_m\}$, the dimension of low dimension space d ;

1. Centralizing to all data points, and calculate the mean of each column and each number subtracts the mean $x_{ij} = x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. x_{ij} is the data of n row and m column. After the centralization, the matrix becomes $\tilde{X}$.
2. Calculating the covariance matrix $\tilde{X}^{\mathrm{T}}\tilde{X}$ of the sample according to the formula;
3. Then the covariance matrix $\tilde{X}^{\mathrm{T}}\tilde{X}$ are being feature decomposition ;
4. Taking the feature vector of corresponding to the number of d largest eigenvalue, and the resulting matrix of feature vector is $W = [w_1, w_2, \cdots, w_d]$ $1 < d < \mathrm{m}$, and each vector of the low dimension matrix W are represented by $w_i = [w_{1i}, w_{2i}, \cdots, w_{mi}]^{\mathrm{T}}$, $i = 1, 2, \cdots, d$.

Output: Low dimensional output $Y = X \times W$.

The traditional PCA algorithm is used to alleviate the "dimensionality disaster" problem, and the PCA algorithm maps m dimension features to the d dimension space $(d < m)$. The PCA algorithm is projected X into Y through a linear transformation matrix W , and the vectors of Y are irrelevant to between each other, so that the information between any two principal components does not overlap.

B. The implementation steps

The algorithm proposed in this paper combines clustering and dimension reduction algorithms, and improve the efficiency by combining the two kinds of algorithm. Specific implementation steps are as follows:

(1) Clustering the variable of m column data, and classify the m column data into k small class according to different standards. Clustering is to bring together data variables with large similarity, and clustering methods used in this paper are k-means and hierarchical clustering.

(2) Reducing the dimension of the data after being clustered into k small class respectively. In this paper, the simulation methods which used to reduce the dimension have PCA, LDA, KPCA, and so on;

(3) Merging the k small class data after being reduced dimensionality, and the number of rows is the same and the number of columns is reduced;

(4) Classifying the merged data by SVM, and the general flow chart is shown in Figure 1.

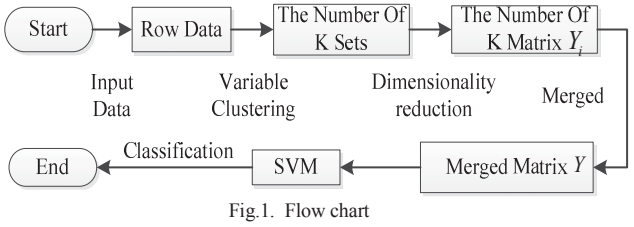

Fig.1. Flow chart

There, it was be a m column of data according to the above algorithms for dimensionality reduction, and the m column of data is represented by $\{X_1, X_2, \cdots, X_m\}$. Variable clustering is performed on the m column data $\{X_1, X_2, \cdots, X_m\}$. After being clustered into k small class, and each cluster is represented by set $C_i, i = 1, 2, \cdots, k$. The first small class is expressed by $C_1 = \begin{Bmatrix} x_{c_1^1}^1 & x_{c_1^1}^2 & \cdots & x_{c_1^1}^{d_1} \\ x_{c_1^2}^1 & x_{c_1^2}^2 & \cdots & x_{c_1^2}^{d_1} \\ \vdots & & & \\ x_{c_1^n}^1 & x_{c_1^n}^2 & \cdots & x_{c_1^n}^{d_1} \end{Bmatrix}$, and the kth small class is expressed by $C_k = \begin{Bmatrix} x_{c_k^1}^1 & x_{c_k^1}^2 & \cdots & x_{c_k^1}^{d_k} \\ x_{c_k^2}^1 & x_{c_k^2}^2 & \cdots & x_{c_k^2}^{d_k} \\ \vdots & & & \\ x_{c_k^n}^1 & x_{c_k^n}^2 & \cdots & x_{c_k^n}^{d_k} \end{Bmatrix}$.Then these k small class data C_i are dimensionally reduced respectively, and the data after being dimensionality reduction are expressed by Y_i respectively. The specific data structure shown in Figure 2.

The total dimension remains unchanged after being clustered, so $\sum_{i=1}^{k} d_i = m$. The dimension after being reduced dimensionality is smaller than the dimension of the data after being clustered corresponding, and $a_i < d_i, i = 1, 2, \cdots, k$. Then merging the k various types of data after being reduced

dimensionality, and the merged matrix is represented by Y. And the number columns in an merged matrix Y is $a=\sum_{i=1}^{k} a_i$, $a < m$,and the number of row has not changed.

C. Dimension r*eduction evaluation method*

The support vector machine (SVM) is a relatively successful classification method. In this paper, the support

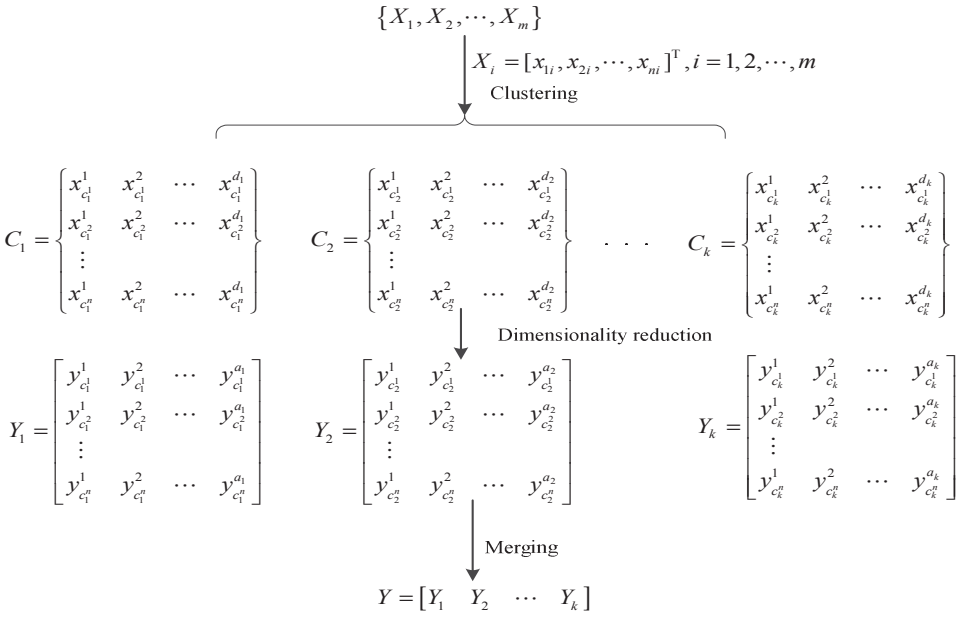

Fig. 2. Data structure diagram

vector machine SVM sorts the data being reduced dimensionality directly or the proposed method. The accuracy of classification is used to judge the effect of dimensionality reduction. The data after being reduced dimensionality as an input data of SVM, and the label of the original data is tagged as the category label of SVM. The number of row after being reduced dimensionality has not changed, and the dimension of column is reduced. Each row still represents a data point, and every column represents a variable. By finding an optimal hyperplane, the distance between the data points and the hyperplane is maximized, and to achieve the effect of two categories. If it is a linear inseparable situation, the SVM needs to map low-dimensional data to high-dimensional space, and the linear discriminant function then divides the feature space into two regions by a hyperplane. Assuming that the data points are represented by x_i , and the data labels are represented by y_i, and $y_i \in \{-1, 1\}$. Finally all data points are divided into two categories. The hyperplane equation is represented by $w_i^{\mathrm{T}} x_i + b = 0, i = 1, 2, \cdots, n$. Assuming the function

$$g(\mathrm{z}) = \frac{1}{1+e^{-z}} \tag{1}$$

Now it substitutes $z = w^T x + b$ for the formula (1), when $h_{\mathrm{w},b}(\mathrm{x}) = g(w^T x + b)$. When the function result $h_{w,b}$ after x_i being put into the equation is equal to zero, it is the corresponding x_i is on the hyperplane. And hope that the model achieves the training goal when $z \geq 0$ $y = 1$; On the contrary, $y = -1$. It is assumed that the functional $h_{w,b}$ make a simplification, and maps it to -1 to 1 simply. .

$$h_{w,b} = \begin{cases} 1, w^{\mathrm{T}} x + b \geq 0 \\ -1, w^{\mathrm{T}} x + b < 0 \end{cases} \tag{2}$$

It needs to be calculated w and b . When new data points appear, they can be sorted directly. How to optimize w and b , it is a key step in supporting vector machine SVM.

Here define the geometric spacing of the points to the hyperplane as margin $\gamma = \frac{w^{\mathrm{T}} x + b}{\|w\|} = \frac{\mathrm{g}(x)}{\|w\|}$, γ is symbolic. What we need here is the absolute value of γ, and need to multiply the corresponding category y , $\tilde{\gamma} = y\gamma = \frac{y(w^{\mathrm{T}} x + b)}{\|w\|} = \frac{y\,\mathrm{g}(x)}{\|w\|}$. The distance between the data points and the hyperplane should be maximized, and find the minimum value from all margins, $\min \gamma_i \quad i = 1, 2, \cdots, n$. Taking into account the geometric spacing of the n data points, and defines the margin as the minimum value for margin of all data points. Finally it is optimize the hyperplane by maximizing the margin. As Figure 3 shown, for example in the two dimension plane, the middle yellow line indicates the optimal hyperplane, and the two lines on both sides are the distance lines. It is need that the interval between two lines is maximized. Because these data points are linearly separable in high

dimension space, so it can use a line to separate the data points. The line represents the optimal hyperplane. And the two lines are just satisfy with the formula $\left|y(w^{\mathrm{T}}x+b)\right|=1$, and the purpose is to make the distance between the two lines as large as possible.

Here it can change the problem of optimize w and b into convex optimization problem by introducing the Lagrangian multiplier α . The Lagrangian function can be used to fuse the constraints into the objective function. The objective function is a formula (3):

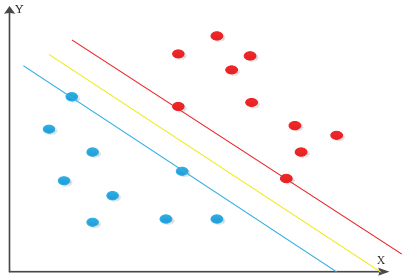

Fig. 3. SVM visual diagram

$$\mathrm{L}(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i(y_i(w^{\mathrm{T}}x_i+\mathrm{b})-1) \qquad (3)$$

Through the training set, finding out the optimal hyperplane and the corresponding parameters w and b . Then the data points can be classified. After all the data points are being classified, then calculate the number of data points that are sorted correctly. The classification accuracy is defined as

$$The\ \ accuracy=\frac{The\ number\ of\ \ data\ correctly\ classified}{The\ number\ of\ \ all\ test\ data} \qquad (4)$$

According to the level of accuracy, and evaluate the effect of the dimensionality reduction. If the accuracy of algorithm proposed in this paper is more high than the method of dimensionality reduction directly, indicating that the method is valid. On the contrary, the efficiency of dimensionality reduction has not improved.

III. NUMERICAL SIMULATION AND ANALYSIS

A. Databases

The data sets used in this paper are Heart scale, Vote and German. Where the first is common test data sets in SVM, and the last two are from the dataset UCI (University of California Irvine). The UCI database is a database for machine learning at the University of California, Irvine. The database currently has a total of 187 data sets. The UCI dataset is a standard test data set commonly. The size of the three datasets are 270×13 , 453×16 , and 1000×24 . The datasets used in this article are tagged data sets for easy testing.

B. Data clustering results

The number of clusters in this algorithm has a great influence on the effect of dimensionality reduction. Because the clustering algorithms used in this paper are k-means and hierarchical clustering algorithm, the initialization center point of k-means algorithm has a great effect on dimensionality reduction. After several simulations, the number of cluster in this paper is taken as two or three in this paper.

C. Evaluation of Data Dimension Reduction

The first two thirds of the general data set are used to train the support vector machine, and one third of the data is used for testing. The data are being dimensionality reduction and the accuracy of classification is calculated, and then calculate the accuracy of classification which the algorithm proposed in this paper. And the two are compared. In the Table I, Table II and Table III below, the first column percentage indicates the accuracy of the classification of the data after being reduced dimensionality directly, and the second column and the fourth column represent the accuracy of combination of different clustering algorithms and dimensionality reduction. The third column and the fifth column represent the increment of the corresponding method. The clustering algorithms used in this paper are k-means and hierarchical clustering algorithm, and the dimensionality reduction algorithms are PCA, KPCA, MDS, LDA, SNE and tSNE. Two kinds of clustering algorithms and six kinds of dimensionality reduction algorithms are matched respectively, and three data sets were analyzed respectively, comparing the accuracy of data directly dimensionality reduction and the method proposed in this paper.

TABLE I. THE ACCURACY OF DATA SETS HEART SCALE

Dimension Reduction		Clustering			
		k-means	Increment	Hierarchical clustering	Increment
PCA	85.0%	86.7%	1.7%	85.0%	0%
KPCA	82.5%	84.2%	1.7%	85.0%	2.5%
MDS	85.0%	85.0%	0%	85.0%	0%
LDA	76.0%	81.0%	5%	81.6%	5.6%
SNE	84.2%	83.3%	-0.9%	84.2%	0%
tSNE	83.3%	82.5%	-0.8%	85.0%	1.7%

As can be seen from Table I, for the data set Heart scale data, the MDS and PCA are better than other dimensionality reduction. The accuracy of the combination of LDA and k-means, hierarchical clustering algorithms are increase 5.6% and 5.0% respectively. Compared with other dimensionality reduction methods, that the methods of combination the clustering and dimensionality reduction are better performance, and the accuracy of the algorithms have been improved in the paper.

As can be seen from Table II, for the data set Vote, the accuracy of the combination of KPCA and hierarchical clustering algorithm has improved 2.8%, and the accuracy of the combination of KPCA and k-means algorithm has improved 3.5%. For this data set Vote, the combination of K-means and KPCA is a superior to the combination of hierarchical clustering and KPCA. In this paper, the accuracy of other algorithms proposed in this paper compared with

other dimensionality reduction algorithms directly, the percentages have been improved.

TABLE II. THE ACCURACY OF DATA SETS VOTE

Dimension Reduction		Clustering			
		k-means	Increment	Hierarchical clustering	Increment
PCA	85.5%	86.9%	1.4%	87.6%	2.1%
KPCA	85.5%	89.0%	3.5%	88.3%	2.8%
MDS	85.5%	87.6%	2.1%	87.6%	2.1%
LDA	82.1%	84.8%	2.7%	84.8%	2.7%
SNE	86.7%	87.7%	1.0%	87.5%	0.8%
tSNE	87.8%	89.6%	1.8%	89.0%	1.2%

As can be seen from Table III, the combination of MDS algorithm and different clustering algorithms have a better effect. Both the percentages of the two algorithms in this paper are increased by 16.9% in this paper. Other the accuracy of the algorithm in this paper compared to other dimensionality reduction directly did have improved.

TABLE III. THE ACCURACY OF DATA SETS GERMAN

Dimension Reduction		Clustering			
		k-means	Increment	Hierarchical clustering	Increment
PCA	70.7%	72.5%	1.8%	72.8%	2.1%
KPCA	70.7%	71.2%	0.5%	71.2%	0.5%
MDS	70.7%	87.6%	16.9%	87.6%	16.9%
LDA	69.2%	72.5%	3.3%	72.8%	3.6%
SNE	70.7%	76.0%	5.3%	76.6%	50.9%
tSNE	70.7%	73.1%	2.4%	73.4%	2.7%

In the course of the experiment, the number of clustering plays a significant role in the effect of classification. After several tests, the number of clustering can't be too large. If the choice is too large, the classification will be wrong. As can see from the three tables, the almost all combinations of clustering algorithms and dimensionality reduction have improved the efficiency for different data sets.

IV. CONCLUSION

In this paper, we proposed a method of combining dimensionality reduction algorithms and clustering algorithms with the aim of improving the accuracy of classification. It is more effective in analyzing the internal structure of data. The methods proposed in this paper have clustering the variables through the similarity of the data firstly, and converts variables into small class with high similarity, and the variables are separated with low similarity. And then the dimension of the data after being clustered was reduced respectively. The data after being reduced dimensionality are being merged. Finally, the merged data are classified by SVM. The validity of the algorithm is judged by the accuracy of classification. In the existing research, the methods of clustering algorithms and dimensionality reduction algorithms are so many. Finding effective combinations from the combination of numerous methods of dimensionality reduction and clustering methods is a research work furtherly, and find the effective combination method for adapting linear and nonlinear data respectively.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No.U1534208， No.61773016, and No. 61703334) and Science and technology plan of Shaanxi Province (No. 2016KJXX-79，and S2015YFJM0027).

REFERENCES

[1] Mnassri B, Adel E M E, Ouladsine M. "Analysis and comparison of an improved unreconstructed variance criterion to other criteria for estimating the dimension of PCA model," Journal of Process Control, 2016, pp.207-223.

[2] Laohakiat S, Phimoltares S, Lursinsap C. "A clustering algorithm for stream data with LDA-based unsupervised localized dimension reduction." Information Sciences, vol. 381, 2017, pp.104-123.

[3] Hettiarachchi R, Peters J F. "Multi-manifold LLE learning in pattern recognition." Pattern Recognition, vol .48, 2015, pp.2947-2960.

[4] Yin X Y, Kong G Y, Zhang G Z. "Seismic attributes optimization based on kernel principal component analysis (KPCA) and application." Oil Geophysical Prospecting, vol.43, 2008, pp.179-183.

[5] Hinton G, Roweis S. Stochastic Neighbor Embedding. "Advances in Neural Information Processing Systems", vol .41, 2002, pp.833--840.

[6] Kokkala J I, Krotov D S, Östergård P R J. "On the Classification of MDS Codes." IEEE Transactions on Information Theory, vol .61, 2015, pp.6485-6492.

[7] Li K C. "Sliced Inverse Regression for Dimension Reduction." Journal of the American Statistical Association, vol .86, 1991, pp.316-327.

[8] Esmin A A A, Coelho R A, Matwin S. "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data." Artificial Intelligence Review, vol .44, 2015, pp.23-45.

[9] Yang Y, Wu Q M J, Wang Y. "Autoencoder With Invertible Functions for Dimension Reduction and Image Reconstruction." IEEE Transactions on Systems Man & Cybernetics Systems, vol .99, 2016, pp.1-15.

[10] Maaten L V D, Postma E O, Herik H J V D. "Dimensionality Reduction: A Comparative Review." Journal of Machine Learning Research, vol .10, 2009, pp.1-34.

[11] Kim H, Howland P, Park H. "Dimension Reduction in Text Classification with Support Vector Machines.". Journal of Machine Learning Research, vol .6, 2005, pp.37-53.`

[12] Cohen M B, Elder S, Musco C, et al. Dimensionality Reduction for k-Means Clustering and Low Rank Approximation[J]. 2015, 46(8):163-172.

[13] Cohen M B, Elder S, Musco C, et al. Dimensionality Reduction for k-Means Clustering and Low Rank Approximation[J]. 2015, 46(8):163-172.

[14] Maaten L V D. "Learning a Parametric Embedding by Preserving Local Structure." Journal of Machine Learning Research, vol .5, 2009, pp.384-391.

[15] Kasun L L, Yang Y, Huang G B, et al. "Dimension Reduction With Extreme Learning Machine." IEEE Transactions on Image Processing, vol .25, 2016, pp.3906-3918.