International Journal of Database Theory and Application
Vol.9, No.1 (2016), pp.87-96
http://dx.doi.org/10.14257/ijdta.2016.9.1.08

Research on High-Dimensional Data Reduction

Cuihua Tian[1, 2], Yan Wang[1], Xueqin Lin[1]*, Jing Lin[3]* and Jiangshui Hong[1]

[1]*school of computer and information engineering, Xiamen University of Technology, Fujian Xiamen 361024, China*
[2]*Universities in Fujian Province Key Laboratory of Things Application Technology, Fujian, Xiamen, 361024, China*
[3]*School of International Languages, Xiamen University of Technology, Xiamen 361024, China*
Tiancuihua@xmut.edu.cn, 2011110702@xmut.edu.cn
343933290@qq.com, 2013110301@xmut.edu.cn, 75154166@qq.com

Abstract

In this paper, features of high-dimensional data are analyzed, and existing problems of the Canonical Correlation Analysis (CCA) are analyzed for a single view of a full supervised view data. In order to improve CCA, we introduce the method of classifier and present a Classifying to Reduce Correlation Dimensionality (CRCD). Meanwhile, combining big interval learning method, we propose the big correlation analysis (BCA). At last, experiments are respectively conducted by using artificial data set and UCI standard data set. The result shows that methods are feasible and effective.

Keywords*: Curse of Dimensionality; Canonical Correlation Analysis ; Classifying to Reduce Correlation Dimensionality; Big Correlation Analysis; Projected Barzilai-Borwein Method*

1. Introduction

Data Mining [1-2] is dedicated to the analysis and understanding of data, revealing hidden data inside knowledge of technology. In recent years, with the emergence of massive data, analysis and processing of large data has become a hot spot of research. These massive data include space remote sensing data, biological data, network data and financial market transaction data and so on. MapReduce [3] is a popular framework for processing large datasets in parallel over a cluster. In contrast, the full range of pathological data analysis and processing is more typical of the Jobs. In order to solve these problems, people should first study the method of high-dimensional data reduction, which is the high-dimensional attribute data how to select and extract from feature [4-10], and mapped to low dimensional data space, so as to discover its intrinsic characteristics, and then make the appropriate and reasonable treatment. Therefore, dimension reduction methods and the related concepts and technology has been put forward and promoted. In order to cater to the trend of the development, this paper to study the relationship between the various dimension reduction from the classical dimension reduction method that is based on dimension reduction methods, and that is canonical correlation analysis (Canonical Correlation Analysis, CCA) [11-12], and according to the data of single view dimension reduction method [13-15] and a series of effective dimension reduction methods. Thus, people will better understand high-dimensional data reduction, effects and methods.

ISSN: 2005-4270 IJDTA

2. High-Dimensional Data Mining

2.1. High-Dimensional Data Mining Features

Data is universal and data mining is an important approach to data analysis and processing. It refers to extract the implicit knowledge and useful information, previously unknown, and potentially technology from a large number of data. High-dimensional data mining is a data mining that based on high dimension, the main difference between the high-dimensional data and the traditional data mining lies in the higher dimensions. At present, the high-dimensional data mining has become a focus and difficulty in data mining. With the development of technology, which makes the collection of data become easier and easier, and results in a bigger and bigger database, more complex, such as the various types trade data, Web documents, gene expression data, document frequency data, rating data of users, WEB usage data and multimedia data, they usually can reach hundreds of dimension attributes thousands of dimension, or even higher.

Due to the existence of high-dimensional data, which makes the research on high-dimensional data mining has a very important significance. But because of the "dimension disaster", it also makes the high-dimensional data mining become extremely difficult, we must adopt some special means to solve. With the increase of the dimension of the data, the performance of high-dimensional index structure decreased rapidly, in the low dimensional space, those who often used Euclidean distance as the similarity measurement, but among data in high-dimensional space, many cases the concept similarity does not exist, it will bring serious test to the high-dimension data mining ,on one hand, it caused the performance of data mining algorithm which based on index structure decreased, on the other hand, a lot of mining method based on distance function will fail. The methods can solve these as the following: by reducing the dimension of the data from high dimension to lower dimension, and then use the approach of the low dimensional data processing to solve; to not decrease the efficiency of the algorithm, We can design more effective index structure, the incremental algorithm and parallel algorithm to improve the algorithm performance of failure; to the problem of failure, we can make it rebirth by redefining. Here is the first method, the problem of dimensionality reduction of high -dimensional data.

2.2. The Data View

The same thing observed from a different perspective will get different information. If we call the result of every view as a view, each view can be obtained with the corresponding data. According to the number of data sources, the data can be divided into single view and multiple view data. A single view of the data refers to the information of observation samples obtained by single channel description, and then this data is only a representation, such as photos, text files, *etc*. Multiple view data refers to the same thing from different angles or in different ways obtained by two or more than two kinds of information description, it is also known as multiple type data, Multiple representation of data, multiple modal data or multiple angle data. For example, the data in the database can be expressed as a multiple view data. This paper focuses on the problem of single view.

2.3. Why the High-Dimensional Data to Dimensionality Reduction?

With the development of science and technology, although the data acquisition of information is more and more abundant, the amount of data becomes more and more huge; the data become more and more complex, more and more data attributes, which brings great pressure to the data analysis and processing. Not only need to pay a high processing power and storage costs, but also led to the emergence of a variety

of problems, such as: due to the distance of any two samples between the high-dimensional are the same, resulting in the failure of the distance measure which based on data mining; the same sample and the increased dimension will make the estimation of sample statistics become more difficult; and the sample also results in "curse of dimensionality" as dimension increases. Therefore, in order to solve these problems with high dimension data, the most directive and effective method is to reduce dimension, that is looking for a low dimensional representation which can reflect the essential attribute of the original data ,and which can reduce the workload and reduce the cost, so that the subsequent work more stable and smooth and efficient.

2.4. The Mature Data Dimension Reduction Method

The methods of the dimensionality reduction include feature selection and feature extraction, and the feature selection is to select the best feature subset to classification from the original high-dimensional attribute according to certain rules, feature extraction is to find a mapping which can transform the original high-dimensional space into a lower dimensional space, and map the original data into this space, which will be scattered the classified information of original features into a small amount new features. They differ in that feature selected by feature selection is still the original feature set, and the new feature extracted from feature is a combination of primitive features, Which often do not belong to the original feature set. Therefore, feature selection is a special form of feature extraction. Dimensionality reduction here only studies from feature extraction. The dimensionality reduction of high-dimensional data has become one of the core issues in the field of data mining. So far, the dimensionality reduction method such as principal component analysis, linear discriminant analysis [16-18], locality preserving projections, CCA, partial least squares regression are relatively mature.

2.5. Problems

The CCA and partial least squares regression are used in the definition of correlation. The relationship between two random variables can be measured by correlation which is of great significance to study the dimensionality reduction. Because the correlation of the CCA and partial least squares regression and its improved form the definition of each are not identical, they measure the correlation between the variables in their own characteristics, which collectively refers to as the generalized correlation. This paper is based on the generalized correlation analysis and carries out the work. Although the generalized correlation analysis not only has got certain development, but also has broad application prospects, the existing research still has a lot of problems and difficulties that are as follows:

(1) In the treatment of single view data, CCA back to the linear discriminant to analyze the effect, which inevitably has the inherent defects of the latter, such as the problem of the small sample size and distribution of the data dependence, the dimension constraint of the dimensionality reduction and so on.

(2)The discriminant information between samples in classification is important, but when the correlation analysis deal with the multiple view data, which only pursuit the maximum of different view data, and the relationship between the size of correlation and classification is not substantial link. This means that, even if we can find the biggest direction of projection from correlation, which makes the correlation reach a maximum in this direction of projection, we are not sure that the sample which is not the same can have a good separability in this direction.

(3)The correlation analysis pay attention to study the corresponding information between data sets from different channels, it requires the target description information / data must be one by one corresponding. However, in practical

application, because of the differences between the sampling frequency, the unexpected failures of the machine, the difficulty of measuring, or the high cost factors, which often leaded to signal is not synchronized, even worse that it missed some corresponding data. so that it discovered a large number of multiple view data of the incomplete pairing set, and formed some half matched data, paired data, weak pair data, then all the dimensionality reduction method which for the one by one pairs of multiple view data can't be used directly.

3. Classifying to Reduce Correlation Dimensionality (CRCD)

3.1. The Study of the CRCD

When dealing with the single view of a full supervised view data, CCA back to the linear discriminant to analysis the effect, the root lies in that the objective function of the CCA used the overall correlation of the maximum samples and the same samples shared the same class label. In order to solve this problem, we should consider the classification, which called Classifying to Reduce Correlation Dimensionality (CRCD).Improved:

First, only according to the individual correlation within class to design objective function, remove the process of the comparison between the correlation, which not only reduces the computational complexity significantly, but also get better performance than correlation discriminant analysis. This method was named as CRCD1.

Second, it is similar with the criterion of the correlation discriminant analysis, and pay attention to the correlation about the within class and between class and class by maximum individual correlation of the within class and between individual and minimize the individual correlation of the between class to construct the objective function. It was named CRCD2.

Third, in the light of the linear separability of most of the data in practice, we use kernel method to nonlinear the above two methods, that is put them in combination with the experience of nuclear, expand the dimensionality reduction of nuclear techniques, the classification performance is improved after dimensionality reduction. They were named EK-CRCD1 and EK-CRCD2.

Specific method: we can use the geometric properties of the regular simplex to design class and according to the class label information of the single view data to construct the dual view data by class label encoding. At the same time, we can combine with the design of classifier and use the individual correlation between the sample and the corresponding class label to design dimensionality reduction method, the purpose of this is to make the maximum correlation between training samples and their corresponding class label and label, and make the minimum correlation with other types of the relationship between class label and label. Thus, we can get two dimensionality reduction methods CRCD1and CRCD2 which based on generalized correlation analysis. Further, which combined with the experience of the nuclear, we also can get the nuclear nonlinear dimensionality reduction method EK-CRCD1 and EK-CRCD2 which based on stronger classification performance.

3.2. The Verification Experiment

Experiments on artificial data set and UCI data set, the results verify the effectiveness of the proposed method for dimensionality reduction. UCI is public test data sets, which provides attributes and categorical data, the user can accord their own data method to UCI data set classification. UCI data set is connected: http://archive.ics.uci.edu/ml/datasets.html

3.2.1. The Dimensionality Dimension Algorithm

```
Input：given a sample  X={x1,x2,x3,……,xn}∈Rp×n
      n is sample size
Import：v=(w,b)
Downdimension{   //Initialization
            Generation of class label encoding   L1,L2,L3,……,LC
               Initialization  V(1)= (w,b)    // C is Classification
                Set the maximum number of iterations   MaxItera
                  Set precision δ,set counter  r=1
                   //Dimensionality reduction
                   While (r<MaxItera &&|V(r)- V(r-1)|< δ)
                   Do {
                   r= r+1
                   Obtain an approximate solution with gradient iteration method V(r)
                         }
               //print result
            Print V = ( W, b )
      }
```

Description: if the dimensionality reduction method is different, and the iterative formula of gradient method is different. Due to limited space formula is complicated, there is no detail.

3.2.2. The Validation Data

The three category of the sample, obey the normal distribution. The data of each class of each of the 100 samples randomly selected from each sample in half for training, the remaining samples for testing. Table 1 shows the 50 average recognition rate of the experiment.

Table 1. The Artificial Data Set Recognition Rate

Method	Co-NN	Ed-NN	CCA	CDA	CRCD1(Te/Tr)	CRCD2(Te/Tr)
Result	0.7987	0.7816	0.7783	0.5838	0.8462/0.8523	0.8632/0.8794

Where: Co-NN: nearest neighbor classifier based on the correlation measure based on Euclidean distance measure; Ed-NN: nearest neighbor classifier; CCA: canonical correlation analysis; CDA: correlation analysis; CRCD1 (Te/Tr) CRCD1 respectively to identify the training set rate in the test set; CCDR2 (Te/Tr) CRCD2 in the test set / recognition rate on the training set respectively.

Table 2 gives the average recognition from the 10 experimental UCI standard test data rate. The Dim/Class/Num dimension of the data, respectively the number of categories and the number of samples.

Table 2. UCI Standard Test Data Set Recognition Rate

Data set(Dim/ Class/Num)	*Recognition rate*							
	Co-N N	Ed-NN	CCA	CDA	CCDR1 (Te/Tr)	EK-CCDR1	CCDR2 (Te/Tr)	EK-CCDR2
Abalone 8/21/4168	0.1922	0.2038	0.1992	0.1885	0.2388	0.2396	0.2347	0.2497
					0.2332	0.236	0.2329	0.2351

						9		
Balance 4/3/625	0.8592	0.8315	0.8796	0.9157	0.8726	0.8860	0.9110	0.9932
					0.8802	0.8864	0.9175	0.9533
Dermatology 33/6/366	0.9667	0.9604	0.9606	0.8523	0.9654	0.9985	0.9773	1.0001
					0.9743	0.9848	0.9778	1.0000
Ecoli 6/6/332	0.8287	0.8212	0.821	0.8113	0.8660	0.8809	0.8948	0.9162
					0.8663	0.8804	0.8939	0.9157
Iris 4/3/150	0.9078	0.9503	0.9562	0.9219	0.9196	0.9382	0.9662	0.9969
					0.9218	0.9375	0.9494	0.9928
Lense 4/3/24	0.7784	0.6158	0.7837	0.5893	0.7927	0.9739	0.8636	0.9891
					0.7873	0.9662	0.8588	0.9931
Soybean 35/4/47	0.9696	0.9718	0.9779	0.9364	0.9753	1.0000	0.9995	1.0000
					0.9939	1.0000	0.9995	1.0001
Teaching 53/8/46	0.5038	0.4673	0.5306	0.4883	0.4240	0.4899	0.5413	0.9724
					0.4184	0.4963	0.5340	0.9615
Waveform 21/3/5000	0.8111	0.7791	0.8140	0.7474	0.8582	0.8625	0.8562	0.8623
					0.8561	0.8592	0.8507	0.8651
Yeast 8/10/1484	0.4626	0.5071	0.5046	0.4908	0.5657	0.5814	0.5569	0.7818
					0.5743	0.5893	0.5658	0.7304

4. Big Correlation Analysis (BCA)

4.1. The Study of the Big Correlation Analysis (BCA)

According to the monitoring of single view data set, combined with interval learning method to design a giant correlation analysis (BCA)dimensionality reduction methods, namely through the maximum projection of samples and their minimum pair correlation of class label to get the large interval analysis method and to achieve supervised dimensionality reduction of single view of data. The method aims to maximize the minimum correlation of all the training samples and the class label, thus overcoming the equivalent defects of the supervised learning and linear judgment and analysis ion the single view data .In fact, the objective function of the method is finally converted to the box constrained quadratic programming problems with two relaxation. Finally, select the part of the data in the UCI data set for experiment, the results verify the validity of BCA.

4.2. The Verification Experiment

4.2.1 The Dimensionality Dimension Algorithm

Input：sample matrix $X=[x_1,x_2,\ldots\ldots,x_n]\in R^{D\times n}$; number of iterations k
Label matrix $Y=[y_1,y_2,\ldots\ldots,y_n]\in R^{(C-1)\times n}$;Parameter δ

Import：$\alpha=(\alpha_1,\alpha_2,\ldots\ldots\alpha_n)^T$,$W=\sum_{i=1}^{n}\alpha_i x_i y_i^T$ //projection matrix W

BCAdowndimension{
Calculate the symmetric positive semidefinite matrix A, elements $A_{ij}= y_i^T y_j x_i^T x_j$, i=1,2,……,n;
Use the PBB method to solve the optimization problem $\min\alpha^T A\alpha/2-l^T\alpha$,
Constraint conditions $o<=\alpha<\eta l$, get $\alpha=\{\alpha_1,\alpha_2,\ldots\ldots\alpha_n,\}^T$
(1) Given $\alpha^{(1)}=\{\alpha_1^{(1)},\alpha_2^{(1)},\ldots\ldots\alpha_n^{(1)}\}^T\in R^n$, $\lambda_1>0$
If $\alpha^{(1)}\notin\Omega$， $\alpha^{(1)}$ replace $P_\Omega(\alpha^{(1)})$， k =1
(2) Calculate the projection vector $g_k=A\alpha^{(k)}$ -1
If $|P_\Omega(\alpha^{(k)}-g_k)-\alpha^{(k)}|_2<\delta$
Stop the cycle and jump to the final output statement
(3) Calculate $\alpha^{(k+1)}=P_\Omega(\alpha^{(k)}-\lambda_k\times g_k)$
(4) Calculate $s_k=\alpha^{(k+1)}-\alpha^{(k)}$, $\lambda_k=s_k^T s_k/(s_k^T(g_{k+1}^T-g_k))$ //long of the step
(5) k = k +1, back (2)
//The following is the final output statements
Import：$\alpha=(\alpha_1,\alpha_2,\ldots\ldots\alpha_n)^T$,$W=\sum_{i=1}\alpha_i x_i y_i^T$

}

Among them: $l=[1,\ldots,1]^T\in R^n$, $o=[0,\ldots 0]^T\in Rn$, Ω as the set of feasible solution of optimization problem is: $\Omega=\{alpha\in R^n|o<=alpha<\eta l\}$, P_Ω is a projection operator from the n-dimensional space to the feasible set.

4.2.2 The Validation Data

19 sets of data from the UCI, contains 13 types of data sets and 6 two types of data sets. For each data set, randomly selected from each sample in half samples for training, the remaining samples for testing, the experiment was repeated 10 times.

Table 3. The Recognition Rate of the UCI Data

Data set (gategory/dimension/number of samples)	CCA	LDA	CDA	CCAs(k)	SVM	BCA
Balance(3/4/625)	87.60±5.17	87.53±4.84	92.24±0.76	89.13±3.12(1)	91.99±1.90	93.94±3.67
Bupa(2/6/345)	57.45±7.79	60.52±14.0	63.27±17.0	60.41±18.74(76)	69.59±6.37	69.65±1.94
Cmc(3/9/1473)	42.61±4.22	42.47±4.33	46.61±3.95	44.95±2.32(24)	51.02±1.59	44.84±1.46
Dermat(6/33/366)	96.64±0.76	96.32±0.47	85.76±4.38	96.48±0.47(2)	97.53±0.88	97.31±0.49
Ecoli(6/6/332)	81.10±5.45	80.36±4.78	80.30±3.39	81.46±7.12(2)	87.93±3.87	81.75±3.50
Glass(6/29/214)	58.57±29.2	54.76±40.0	58.01±8.97	60.28±6.54(4)	64.29±12.0	65.33±4.27
Iris(3/4/150)	92.80±4.80	93.33±3.	91.33±9.1	93.60±3.48(1	97.87±3.6	97.07±1.

		56	9	1)	4	50
Ionosphere(2/34/351)	82.40±6.30	84.29±5.09	83.43±4.14	83.14±6.62(3)	85.94±8.13	89.83±4.34
Lense(3/4/24)	79.09±50.0	79.08±56.0	57.27±221	78.18±77.14(2)	78.18±77.0	85.44±44.2
Pid(2/8/768)	68.89±4.12	68.89±4.12	66.09±3.81	70.44±5.93(156)	77.81±1.26	65.91±4.69
Sonar(2/60/208)	75.92±29.0	74.56±23.0	80.39±12.0	77.48±35.78(29)	79.13±15	83.69±9.80
Soybean(4/35/47)	97.38±5.04	97.39±5.04	92.17±46.0	97.83±5.25(2)	100±0.0	99.13±7.57
Teaching(3/5/151)	55.07±38.0	55.60±79.0	50.67±23.0	53.86±40.38(20)	53.47±22.0	58.40±20.0
Thyroid(3/5/215)	92.71±5.20	93.27±7.14	78.32±13.0	93.46±3.88(9)	95.17±4.55	96.73±1.79
Vehicle(4/18/846)	73.44±4.82	73.85±3.05	58.01±2.71	73.01±2.65(9)	79.03±1.92	52.86±5.36
Wine(3/13/178)	98.08±0.87	97.84±1.56	63.64±16.0	97.38±3.76(5)	96.45±5.29	79.89±22.0
Water(2/38/116)	70.87±51.0	71.40±32.0	84.91±9.72	74.09±14.92(4)	78.07±17.0	86.67±14.0
Wdbc(2/30/569)	93.45±2.37	93.44±2.38	90.25±6.11	95.08±1.11(53)	94.40±1.97	92.01±2.84
Waveform(3/21/5000)	81.23±0.48	81.22±7.96	74.64±0.24	81.84±0.39(170)	86.49±0.21	82.01±0.18
Average	78.17±13.39	78.22±15.7	73.54±21.33	79.05±12.61	82.36±13.27	80.12±12.1

Among them, the canonical correlation analysis; CCA: LDA: CDA: linear discriminant analysis; correlation analysis; CCAs (k): soft label: SVM CCA; support vector machine [15-16].

5. Conclusion

Dimensionality reduction is an important research direction in the field of statistical pattern recognition, and the results of the dimensionality reduction directly affect the final performance of the overall pattern recognition system. At present, the dimensionality reduction method based on correlation is an important branch of data dimensionality reduction, and has attracted more and more attention, and further become a hot research topic in pattern recognition learning, many areas of the scope of its application to image analysis and image processing, multimedia processing, medicine, marine meteorological forecast data analysis, computer visual information retrieval and cross language text classification.

This paper starts from the classical dimensionality reduction methods which are based on the analysis of the relationship, in light of the single view data, and analysis the reason which is equivalent to linear discriminant analysis in the whole supervision situation, and design a new series of dimensionality reduction methods to overcome this difficulty. Classification of dimensionality reduction introduce the classifier in the canonical correlation dimensionality reduction, the purpose of this is to make the maximum correlation between training samples and their corresponding class label and label, and make the minimum correlation with other types of the relationship between class label and label, and use the individual correlation between the sample and the corresponding class label to design dimensionality reduction method. Giant correlation analysis dimension reduction method is inspired by the large margin learning method, the maximum minimum correlation of all the training samples and the class label, in order to

overcome the drawbacks of the linear equivalent supervised learning with a single view of the data on the judgment, to reduce the dimension of high-dimensional data. Finally, the effectiveness of the proposed dimensionality reduction methods are verified by experiment. The next step, also need to consider the problem of multiple view data dimension reduction.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China No. 61373147; Science and technology research projects of XMUT No.YKJ11012R and No.YKJ10037R; and the 12th five-year plan project of Xiamen education science No.1250.

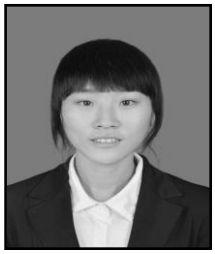
Reference

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd ed. Morgan Kaufmann Publishers, **(2006)**.
[2] P. N. Tan and M. Steinbach, "Introduction to Data Mining, Posts & Telecom Press", Beijing, **(2011)**.
[3] X. Tang, L. Wang and Z. Geng, "A Reduce Task Scheduler for MapReduce with Minimum transmission Cost Based on Sampling Evaluation", International Journal of Database Theory and Application, vol. 8, no. 1, **(2015)**, pp.1-10.
[4] Mitra P., Murthy C. A. and Pal S. K., "Unsupervised Feature Selection Using Feature Similarity", IEEE Trans on Pattern Recognition and Machine Intelligence, vol. 3, no. 24, **(2002)**.
[5] Ran G. B., Amir N. and Naftali T., "Margin Based Feature Selection-theory and Algorithms", Proceedings of the 21th International Conference on Machine Learning, Banff, Canada, June: 43-50, **(2004)**.
[6] Cang S. and Yu H. N., "A New Approach for Detecting the Best Feature Set. Proc of Networking", Sensing and Control. [S. l.] : IEEE CNF, September **(2005)**, pp. 74-279.
[7] Peng H. C., Long F. H. and Ding C., "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 27, **(2005)**, pp. 1226-1238.
[8] Gheyas I. A. and Smith L. S., "Feature Subset Selection in Large Dimensionality Domains", Pattern Recognition, vol. 43, **(2010)**, pp. 5-13.
[9] Byeon B. and Rasheed K., "Selection of Classifier and Feature Selection Method for Microarray Data", 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), Washington, DC, December **(2010)**.
[10] Santos J. M. and Ramos S., "Using a Clustering Similarity Measure for Feature Selection in High Dimensional Data Sets", Proceedings of ISDA' 2010, Cairo, November **(2010)**.
[11] J. Leps and P. Smilauer, "Multivariate Analysis of Ecological Data using CANOCO", Cambridge University Press, **(2003)**.
[12] Qin S. J., "Recursive PLS Algorithms for Adaptive Data Modeling", Computers & Engineering, vol. 4-5, no. 22, **(1998)**, pp. 503-514.
[13] Wasikowski M. and Chen X. W., "Combating the Small Sample Class Imbalance Problem Using Feature Selection", IEEE Transaction on Knowledge and Data Engineering, vol. 10, no. 22, **(2010)**, pp. 1388-1400.
[14] J. P. Van der Maaten, E. O. Postma and H. J. V. Anden Herik, "Dimensionality Reduction: A Comparative Review", Preprint submitted to Journal of Machine Learning Research, **(2009)**.
[15] L. J. P. V. d. Maaten, E. O. Postma and H. J. V. D. Herik, "Dimensionality Reduction: a Comparative Review", Tilburg University, **(2008)**.
[16] Roweis S. and Saul L., "Nonlinear Dimensionality Reduction by Locally Linear Embedding", Science, vol. 290, **(2000)**, pp. 2323- 2326.
[17] Cristianini N. and Shawe-T. J., "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, **(2000)**.
[18] Liu Y. and Zheng Y. F., "FS-SFS: A Novel Feature Selection Method for Support Vector Machines", Pattern Recognition, vol. 39, **(2006)**, pp. 1333-1345.

Authors

Cuihua Tian, She was born in 1970, an associate Professor and graduate Supervisor of the school of Computer & Information Engineering, Xiamen University of Technology. She he received her Ph.D. of education in Northeastern University. Her research interests includes Algorithm, Network, the Internet of Things, Intelligent Information Processing, Data Mining, Big data.

Yan Wang, He was born in 1977, a University Lecturer of School of Computer & Information Engineering, Xiamen University of Technology. He got a Ph D. in the school of computer and information engineering, Renmin University of China. He is mainly researching on Database, Data Mining and Big Data.



Xueqin Lin, She was born in 1991, a student of the school of computer and information engineering, Xiamen University of Technology. Her major is software service engineering. In daily life, she reads newspapers and books related computer in many aspects, and study the computer knowledge of the frontier seriously. She is interested in big data, data mining, and software engineering.

Jing Lin, She was born in 1988, a Teaching Assistant of School of International Languages, Xiamen University of Technology. She got a Master's degree in Accounting at State University of New York at Albany. She is mainly researching on Business English.

Jingshui Hong, born in 1992, a student of School of computer and information engineering, Xiamen University of Technology. He is mainly researching on computer graphics / image algorithm, 3D reconstruction algorithm, and Data Mining *etc*.