International Journal of Database and Theory and Application
Vol.9, No.2 (2016), pp. 109-118
http://dx.doi.org/10.14257/ijdta.2016.9.2.12

An Effective SVM Ensemble Algorithm Based on Different Thresholds of PCA

Yukai Yao[1,a], Bo Wang[2], Qingjun Yang[2,3], Dongsheng Ji[2], Tao Ma[2] ,* and Xiaoyun Chen[2,b]

[1]*College of Computer and Communication, Lanzhou University of Technology, 247 Lan'gongping Road, Lanzhou 730050, China;*
[2]*School of Information Science & Engineering, Lanzhou University, 222 South Tianshui Road, Lanzhou 730000, China;*
[3]*Qinghai Province Meteorological Bureau, 19 Wusi Road, Xi'ning, China.*
[a]*yaoyukai@163.com;* [b]*chenxy@lzu.edu.cn*

Abstract

This paper proposes an effective ensemble classifier, named PCAenSVM, which consists of ten weak Support Vector Machine classifiers based on different Principal Component Analysis thresholds. Those ten base Support Vector Machine classifiers are made up to fulfill classification tasks using Majority Voting strategy. Experiments are made on four UCI data sets and a data set from the Uppsala University to evaluate the performances of PCAenSVM. The results of PCAenSVM are compared with that of LibSVM and EnsembleSVM. Experimental results show that PCAenSVM has better classification accuracy than other two algorithms. Moreover, PCAenSVM has the same confidence level with the LibSVM, and its confidences of accuracy and sensitivity on those five data sets outperform that of the EnsembleSVM.

Keywords*: Support Vector Machine, Ensemble Methods, Principal Component Analysis, Majority Voting, Classification*

1. Introduction

Support Vector Machine (SVM) is a hopeful classification and regression algorithm, which is based on the VC Dimension theory and the Structural Risk Minimization theory, both of which are the core contents of the Statistical learning Theory. SVM was first presented by Cortes and Vapnik in 1995 [1], and it has been applied successfully in text categorization, bioinformatics, speech recognition and other domains.

The study and application of SVM has been one of the focuses in the field of machine learning for more than ten years. Least squares SVM was proposed by J.A.K. suykens and J. Vandewalle in 1999, a set of linear equations has to be solved instead of a quadratic programming problem in this method. The least squares SVM can be used in more classification problems [2]. Weighted SVM achieved an excellent result on balanced data sets [3]. Active learning with SVM was proposed by Simon Tong in 2002, the method was effective on small datasets [4]. Classifier ensemble is one direction of machine learning research, and an ensemble of classifiers can improve the performance of an individual classifier [5]. The ensemble method was introduced to SVM in 2002, and, it outperformed a single SVM in terms of accuracy [6]. A data driven ensemble classifier was used in credit scoring system by Hsieh and Hung in 2010, and it showed significantly better performance than that of the conventional ensemble classifiers [7]. A KNN-SVM ensemble classifier was applied to predict the subcellular localization of eukaryotic proteins, and a good performance of predictions was achieved [8]. A SVM ensemble was

ISSN: 1975-0080 IJDTA

used in steganalysis of digital media, which reduced the training complexity [9]. Recently, Ensemble classifier has been a research focus of machine learning. Base (or weak) classifiers are important for an ensemble classifier. One method is using the same classification algorithms to aggregate the ensemble classifier, the other method is to use different classification algorithms to form an ensemble classifier. LibSVM is a SVM tool, which is implemented by Professor Chih-Jen Lin in the National Taiwan University [15]. EnsembleSVM is a free software package, which uses Support Vector Machines (SVM_s) as base models to achieve ensemble learning [16].

Although lots of improvements of the adapting and using for SVM have been made, most of them are limited to solve the special problems or tasks, and the overall performances of them are still undesirable. In this paper, we adopt the first ensemble strategy mentioned above to construct the SVM ensemble: different base SVM classifiers are based on different PCA thresholds, and ten PCA based SVM classifiers are aggregated using majority voting method. Diversity is necessary for obtaining an accurate ensemble [10-12], and it inspired our idea of presenting PCAenSVM. PCA is the tool which is used to produce the diversity among base classifiers in this paper. PCA is a preprocessing method which can reduce redundancy of high-dimensional data while retaining the important information of data. Retaining different principal components, several new datasets will be created. Training on such new datasets independently, several base classifiers can be produced. Then, these base classifiers are grouped together using the majority voting strategy, thus an ensemble classifier is produced, which can be used to classify the test data.

The performance of the proposed ensemble algorithm PCAenSVM is evaluated on five different datasets, and the experimental results show that PCAenSVM can achieve a higher accuracy than LibSVM and the EnsembleSVM, while spending a little more time than LibSVM and EnsembleSVM.

2. Principal Component Analysis

PCA is a technique that can reduce or eliminate redundancy of high-dimensional data [17]. Using linear transformation, PCA can reduce the dimensionality of data while retaining most of the information in the datasets.

Given training set X, which has N data, a D-dimensional attribute, it may create a matrix of N*D. Using PCA technique, the original data space is mapped into a M-dimensional subspace, the data Xpca is generated which contains the same information with X, while X_{pca} is a matrix of N*M, Where M$\leqslant$D. The low letter u is unit vector in M-dimension of training set after mapped. Let $\overline{X}$ be the mean vector of training set X, it can be calculated using Eq.1.

$$\overline{X} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1}$$

Then the variance of data after mapped is Eq. 2

$$\frac{1}{N}\sum_{n=1}^{N}\left\{u^T x_n - u^T \overline{x}\right\}^2 = u^T S\ u \tag{2}$$

Where S is the covariance of training set X. The goal is to maximize the covariance Eq. 2. Because u vectors are orthogonal vectors, Lagrange multiplier method can be introduced to solve the optimization problem, then a constrained equation is constructed as Eq. 3

$$u^T s u + \lambda(1 - u^T u) \tag{3}$$

By differentiating on u, the maximum value can be obtained, letting the differential coefficient equal to zero, Eq. 4 can be obtained.

$$Su = \lambda u \tag{4}$$

Where λwill be for the eigenvalues of the covariance matrix S, and u is its corresponding eigenvector. Then Xpca can be calculated as Eq. 5

$$X_{pca} = Su \tag{5}$$

Where X_pca is a matrix of N*M which can be the new training set instead of X, thus reducing the number dimensions and computational cost. Given different thresholds, after PCA processing, different X_{pca} which represents different number of principal components are obtained. PCA is performed with a threshold ranging from 10% to 100%, and ten new training sets are generated in this paper.

3. Support Vector Machine Classification

In this section, a brief of the SVM classifier is provided. A simple SVM classifier is used to separate the data set into two classes.

As shown in Figure 1, each black or small white dot is a data (xi,yi), where x_i∈R^N, y_i∈{-1,1}. The function g(x)=w·x+b is a hyperplanes, then g(x)=-1 and g(x)=1 are two hyperplanes paralleling to g(x)=0. There are many hyperplane which can separate these data into two class. The goal is to maximize the margin between g(x)=0 and g(x)=1. The hyperplane g(x)=0 is called optimal separating hyperplane (OSH) and the data on OSH are called support vectors [18]. The distance between data xi and the hyperplane g(x)=0 is calculated as d=|g(x_i)|/ ‖ w ‖ . Maximizing the margin is equivalent to maximizing ‖ w ‖ ^{2}. Thus, the problem to find optimal separating hyperplane is summarized as Eq. 6.

$$\min \frac{1}{2}\|w\|^2 \quad s.t. \quad y_i\left[(w \cdot x_i) + b\right] - 1 \geq 0 \tag{6}$$

Eq. 6 is a constrained minimization problem and it can be solved using Lagrange multipliers. $f(x =) \sum_{i \in R} \alpha \hat{s}_i^* y \quad x \; g x_i^* + b \; n$ (7)

Eq. 7 is the final decision function of SVM. Given a test data xt, which class the data belongs to depends on the decision functionf(xt).

In most cases, as showed in Figure 2, training set does't be separated by a hyperplane, kernel functions is introduced to solve this problem. Kernel functions has been used to mapped data set into a higher dimension space [19]. Therefore, the mapped data set can be separated by a OSH. Then decision function is calculated according to Eq. 8.

$$f(x) = \mathrm{sgn}\{\sum_{i \in R} \alpha_i^* y_i^* K(x_i, x) + b^*\} \tag{8}$$

There are four frequently-used kernel functions which are used to map low-dimension data space into higher-dimension data space. They are linear, polynomial, radial basis, and sigmoid [20]. Radial Basis Function (RBF) kernel function is used in this paper. A SVM tool named LibSVM is used to make binary classification.

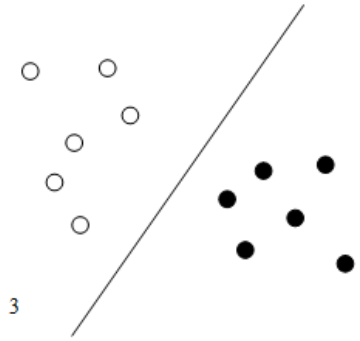

Figure 1. Linearly Separable

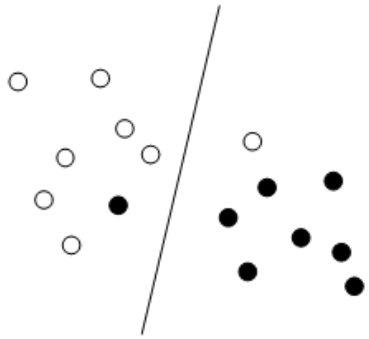

Figure 2. Linear Inseparable

4. Ensemble of Classifiers

Ensemble is a statistical and machine learning method. An ensemble classifier consisted by a group of several base classifiers. These independent base classifiers' classification results on test datasets are combined in some strategy to make a collective decision. An ensemble of classifiers often give a better performance than individual classifiers [13]. Majority voting, least squares estimation based weighting, and double-layer hierarchical combination are often used to combine the base results to make the final decision [6]. Majority voting is adopted in this paper. Diversity is the key to produce an accurate ensemble classifier [10-12]. Training on different datasets will produce different base classifiers. And how to generate new training datasets by sampling the training set is important. Both bagging and Adaboost select samples from the original training dataset with replacement randomly. They allow the original samples to be repeated in the new training dataset, while others may be left out [14]. Bagging and adaboost are two classical meta-algorithms to aggregate an ensemble classifier.

5. Ensemble SVC on different thresholds of PCA

Bagging [21], AdaBoost [22] and Random Forest [23] are used generally to sample a diversity of training sets from original data, then a series of classifiers on these new training sets are trained independently. These new produced classifiers can be combined together using majority voting or other methods to make a final decision on testing set. In order to produce a diversity of datasets from original datasets, principal component analysis (PCA) is utilized in this paper. A PCA process is equivalent to a sampling in bagging or adaboost. Given ten different threshold from ten percent to one hundred percent, ten different training sets can be created. The optimal separating hyperplane of

these ten training sets will also differs from each other. These new datasets contain different amount of principal components, ranging from ten percent to a hundred percent. This idea is the source of PCAenSVM algorithm. Algorithm 1 shows the pseudo code of the PCAenSVM.

Algorithm 1: PCAenSVM

Input: original instances x_{raw}
Output: predicted label and classification accuracy

1: divide the original data x_{raw} into testing dataset x_t, y_t and training set x_{tr}, y_{tr}, using random sampling;
2: for i = 1 →10 do
3: threshold←i*10/100
4: perform PCA on on both x_{tr} and x_t, output is x_{trpca} and x_{tpca}
5: use SVM to produce a model m_i and predicted label $Plabel_i$
6: end for
7: perform majority voting on $Plabel_i$ and ouput the last label Plabel and the classification accuracy

Random sampling is used to select testing set. The testing set are about a third of the original data set in our experiments. Maintaining ten different amount of principal components, ten different training sets are generated from original training sets. Each new training set is smaller than original training set but maintain most information of original training set. On these new training sets, ten different SVM classifiers will be obtained. These ten SVM classifiers can be combined to classify the testing set using the strategy of majority voting. The Figure 3 presents the framework of the ensemble SVM classifier using majority voting method.

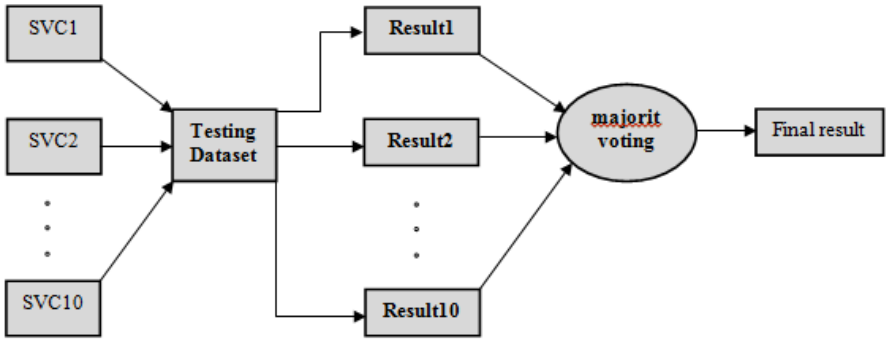

Figure 3. The Framework of PCAenSVM

6. Experiments and Analysis

The performance of PCAenSVM is compared with that of the LibSVM and an ensemble SVM algorithm named EnsembleSVM. Experiments are done on five different data sets. RBF kernel function is chosen in the experiments. Parameter c is set to 1 and gama is set to the default value which equal to one of the number of features of data. The program is implemented in Matlab and run on an machine has the configuration of Intel Core 2 Duo with 2.53GHz CPU and 3GB RAM.

Four data sets from the UCI machine learning repository and another data set from Jan Conrad of Uppsala University are adopted as our experimental data, all of which are preprocessed by Chih-Jen Lin [24]. Training set is about two thirds of the original data set, and the rest one thirds is used as testing set. These five data sets are described in Table 1. In order to make binary classification, data sets are processed. Splice has 3 classes which include EI, IE and Neigher. EI data and IE data are merged into one class and both of these two kinds of DNA sequence are regard as splice position. There are also 3 classes in datasets Cardiotocography, including normal, suspect and pathologic. Suspect data and pathologic data are regarded as one class in this paper, which are set to the label -1, and the normal data are set to 1.

Table 1. Data Sets Description

DataSets Name	Source	Training size	Testing size	Feature size
Cardiotocography	UCI	1500	626	22
Splice	UCI	1186	1400	180
Ionosphere	UCI	240	111	34
Sonar	UCI	140	68	60
Svmguide1	NTU	2030	1059	4

The experimental results of three classification methods are showed and compared in this section. The performance of PCAenSVM is evaluated from the aspects of accuracy, sensitivity runtime and confidence. Each method is performed 5 times and the average values are shown in Table 2, Table 3 and Table 4, respectively. Table 2 shows the accuracy of these three methods, PCAenSVM has a slightly higher classification accuracy than LibSVM and EnsembleSVM, only on dataset Cardiotocography, the accuracy of PCAenSVM is lower than EnsembleSVM. Table 3 shows the classification sensitivity on the five data sets. On two datasets PCAenSVM has a higher classification sensitivity than other two method. On another datasets Ionosphere, PCAenSVM has the same sensitivity with LibSVM, both of their sensitivity is higher than EnsvmbleSVM. On other two datasets, PCAenSVM takes the second position on performance of sensitivity. The running time of this three algorithm are showed in Table 4. The run time of PCAenSVM are longer than that of LibSVM. PCAenSVM outperform EnsembleSVM on two datasets but on other three datasets PCAenSVM has a longer time than Ensemble so that it is difficult to find which algorithm is better on time cost.

Table 2. Classification Accuracy

Name	LibSVM	**EnsembleSVM**	PCAenSVM
Cardiotocography	80.5112	82.97126	82.2684
Splice	92.5801	88.49916	**93.2546**
Ionosphere	92.7928	92.7928	**93.6937**
Sonar	80.8824	82.05882	**82.3529**
Svmguide1	74.5987	91.08594	**94.2398**

Table 3. Classification Sensitivity

Name	LibSVM	EnsembleSVM	PCAenSVM
Cardiotocography	79.9669	99.5868	95.0413
Splice	92.7959	85.7633	**93.1389**
Ionosphere	98.75	93.75	**98.75**
Sonar	80.6452	80.6452	**83.871**
Svmguide1	99.8516	89.911	95.8457

Table 4. Run Time

Name	**LibSVM**	**EnsembleSVM**	**PCAenSVM**
Cardiotocography	0.872989	1.3322	6.3588972
Splice	1.499442	2.0528	7.0322442
Ionosphere	0.010774	0.9654	**0.3092782**
Sonar	0.008703	0.5706	**0.2157228**
Svmguide1	1.340879	1.0016	10.3080828

The error bar plots of classification accuracy of the three algorithms are presented in Figure 4. In the figure, the X-axis represents the five different datasets. Each algorithm is performed 5 times. For each algorithm on each dataset, the mean value of this five results are used by this figure. As per the error bar plot, the PCAenSVM classifier and the LibSVM classifier give results of higher confidence than that using EnsembleSVM. While the PCAenSVM classifier has a higher accuracy than LibSVM and Ensemble generally. The result of sensitivities of three method are plotted in Figure 5. As the figure presents, as similar to the result of accuracy presented in Figure, the variance of sensitivity is zero and has an excellent confidence in sensitivity. The sensitivity of PCAenSVM are higher than other two algorithms generally.

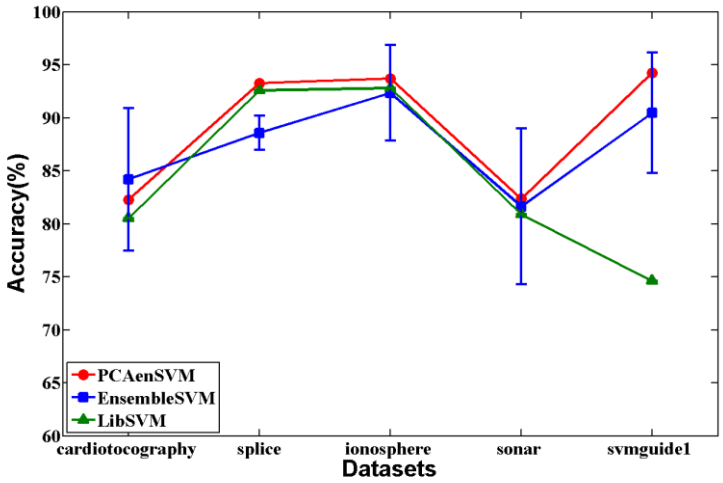

Figure 4. Accuracy of the Three Classification Algorithm

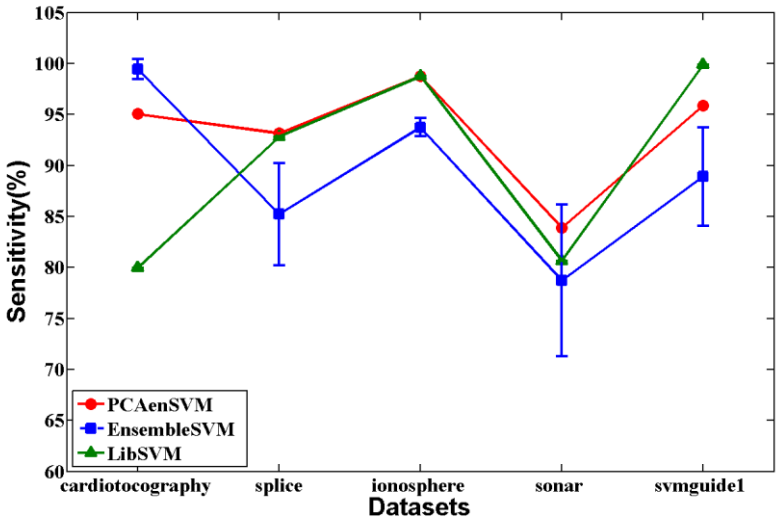

Figure 5. Sensitivity of the Three Classification Algorithm

7. Conclusion

In this study, we present a SVM ensemble algorithm named PCAenSVM, in which, PCA of different threshold values are used to create new training sets. Ten base SVM classifiers which are trained on ten new training sets are aggregated to classify testing sets. PCAenSVM is evaluated on five different datasets, results of experiments show that PCAenSVM achieves a high level of accuracy in binary classifications than that of LibSVM and EnsembleSVM. The PCAenSVM classifier also has a higher confidence than EnsembleSVM. In this paper, ten base SVM classifiers are simply grouped without any selection, in the future, we will select some best base classifiers to produce a more efficient ensemble classifier, so as to improve the classification accuracy and reduce the computational cost further.

Acknowledgement

This study is Supported by the Fundamental Research Funds for the Central Universities (lzujbky-2013-229, lzujbky-2014-47). We thank our tutor professor Xiaoyun Chen for her help and support during this research.

References

[1] V. Vapnik, "The nature of statistical learning theory", Springer, **(2000)**.
[2] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers", Neural processing letters, vol. 9, no. 3, **(1999)**, pp. 293-300.
[3] C. F. Lin and S. D. Wang, "Fuzzy support vector machines", Neural Networks, IEEE Transactions on Neural Networks, vol. 13, no. 2, **(2002)**, pp. 464-471.
[4] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification", The Journal of Machine Learning Research, vol. 2, **(2002)**, pp. 45-66.
[5] T. G. Dietterich, "Machine learning research: Four current directions", The AI Magazine, vol. 18, no. 4, **(1998)**, pp. 97-136.
[6] H. C. Kim, S. Pang and H. M. Je, "Pattern classification using support vector machine ensemble", Proceedings of 16th International Conference on Pattern Recognition, **(2002)**.
[7] N. C. Hsieh and L. P. Hung, "A data driven ensemble classifier for credit scoring analysis", Expert Systems with Applications, vol. 37, no. 1, **(2010)**, pp. 534-545.
[8] L. Q. Li, H. Kuang and Y. Zhang, "Prediction of eukaryotic protein subcellular multi-localization with a combined KNN-SVM ensemble classifier", Journal of Computational Biology and Bioinformatics Research, vol. 3, no. 2, **(2011)**, pp. 15-24.

[9] J. Kodovsky, J. Fridrich and V. Holub, "Ensemble classifiers for steganalysis of digital media", Information Forensics and Security, IEEE Transactions on Information Forensics and Security, vol. 7, no. 2, **(2012)**, pp. 432-444.
[10] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning", Advances in neural information processing systems, **(1995)**, pp. 231-238.
[11] L. I. Kuncheva, "Combining pattern classifiers: methods and algorithms", John Wiley & Sons, **(2004)**.
[12] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy", Machine learning, vol. 51, no. 2, **(2003)**, pp. 181-207.
[13] L. K. Hansen and P. Salamon, "Neural network ensembles", IEEE transactions on pattern analysis and machine intelligence, vol. 12, no. 10, **(1990)**, pp. 993-1001.
[14] Z. Yu, Z. Deng and H. S. Wong, "Identifying protein-kinase-specific phosphorylation sites based on the bagging–adaboost ensemble approach", IEEE Transactions on Nano-Bioscience, vol. 9, no. 2, **(2010)**, pp. 132-143.
[15] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2 no. 3, **(2011)**, pp. 27.
[16] M. laesen, F. D. Smet and J. A. K. Suykens, "Ensemble SVM: A library for ensemble learning using support vector machines", The Journal of Machine Learning Research, vol. 15, no. 1, **(2014)**, pp. 141-145.
[17] A. D. C. Chan and G. C. Green, "Myoelectric control development toolbox", Proceedings of 30th Conference of the Canadian Medical & Biological Engineering Society, **(2007)**.
[18] Q. Wu and D. X. Zhou, "SVM soft margin classifiers: linear programming versus quadratic programming", Neural computation, vol. 17, no. 5, **(2005)**, pp. 1160-1187.
[19] S. Chaplot, L. M. Patnaik and N. R. Jagannathan, "Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network", Biomedical Signal Processing and Control, vol. 1, no. 1, **(2006)**, pp. 86-92.
[20] C. Xu, F. Dai and X. Xu, "GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China", Geomorphology, vol. 145, **(2012)**, pp. 70-80.
[21] G. Zararsiz, F. Elmali and A. Ozturk, "Bagging Support Vector Machines for Leukemia Classification", development, vol. 1, **(2012)**, pp. 2.
[22] S. K. Mathanker, P. R. Weckler and T. J. Bowser, "AdaBoost classifiers for pecan defect classification", Computers and Electronics in Agriculture, vol. 77, no. 1, **(2011)**, pp. 60-68.
[23] C. H. Hsieh, R. H. Lu and N. H. Lee, "Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks", Surgery, vol. 149, no. 1, **(2011)**, pp. 87-93.
[24] Information, http://www.csie.ntu.edu.tw/~cjlin/LibSVMtools/datasets/ binary.html# svmguide1 .

Authors

Yukai Yao, He received the BS degree from the Department of Computer Science and Technology, Northwest Normal University, in 1997, and the Ma degree from School of Information Science and Engineering, Lanzhou University, in 2011. He is now a doctoral student in the School of Information Science and Engineering, Lanzhou University. His research interests include high performance computing, pattern recognition and data mining. He is a member of CCF and IET.

Qingjun Yang, He graduated from the University of Electronic Science and Technology of China in 2008, and received his BS degree. He received his Ma degree from Chengdu University of Information Technology in 1994. He is now a doctoral student in the School of Information Science and Engineering, Lanzhou University, and he is working with the Qinghai Province Meteorological bureau. His research interests include data mining and meteorological data analysis.

Dongsheng Ji, He received the BS degree from School of Information Science and Engineering, Lanzhou University, in 2007, and the Ma degree from School of Computer and Communication, Lanzhou University of Technology, in 2011. He is now a doctoral student in the School of Information Science and Engineering, Lanzhou University. His research interests include biomedical image analysis, pattern recognition and data mining.

Tao Ma, He has being studying the Ph.D. degree in School of Information Science and Engineering, Lanzhou University, Lanzhou, China, from 2013. Since 2003, he is a lecturer at School of Mathematical and Computer Science of Ningxia Teachers University. His research interests include data mining, artificial intelligence and pattern classification.

Xiaoyun Chen, She received the BS degree from the Department of Computer Technology, Jilin University, and the MA degree in the Institute of Atomic Energy, Chinese Academy of Sciences, in 1995. Professor, PhD supervisor, the Director of Institute of Computer Software and Theory, the Director of national Linux Technical Training and Promotion Center of Lanzhou University, the Director of IBM Technology Center of Lanzhou University, Senior Member of CCF, the member of Database Special Committee, the member of Theoretical Computer Science Special Committee, the member of Liberal Arts Computer Basic Teaching Instruction Committee of the Ministry of Education.

The main research fields are data warehouse design and construction, data mining algorithms and applications, special data mining, high performance computing, parallel data mining, search engine technology and method, weather information processing, big data, *etc.*