Non-Negative Matrix Factorization and Support Vector Data Description Based One Class Classification

Liyong Ma[1], Naizhang Feng[2] and Qi Wang[3]

**[1] School of Information & Electrical Engineering, Harbin Institute of Technology at Weihai
Weihai, 264209, China**

**[2] School of Information & Electrical Engineering, Harbin Institute of Technology at Weihai
Weihai, 264209, China**

**[3] Department of Automatic Measurement and Control, Harbin Institute of Technology
Harbin, 150001, China**

Abstract

One class classification is widely used in many applications. Only one target class is well characterized by instances in the training data in one class classification, and no instance is available for other non-target classes, or few instances are present and they cannot form statistically representative samples for the negative concept. A two-step paradigm employing non-negative matrix factorization (NMF) and support vector data description (SVDD) for one class classification training of non-negative data is developed. Firstly, a projected gradient based NMF method is used to find the hiding structure from the training instances and the training instances are projected into a new feature space. Secondly, SVDD is employed to perform one class classification training with the projected feature data. Classification examples demonstrate that the proposed method is superior to principal component analysis (PCA) based SVDD method and other standard one class classifiers.

Keywords: *Non-Negative Matrix Factorization, Support Vector Data Description, One Class Classification.*

1. Introduction

In recent years, there has been considerable interest in one class classifiers. One class classification was originally proposed in object recognition application [1]. Only one target class is well characterized by instances in the training data in one class classification problems, and no instance is available for other non-target classes, or few instances are present and they cannot form statistically representative samples for the negative concept [2-4]. This is true in many classification applications, such as fault diagnosis and object identification [1, 4]. In fault diagnosis, it is easy to obtain the normal operation instances, but we have few instances or no instance to model the fault class. In object identification, it is very challenging to collect non-object samples when we train the machine to learn an object, because too many samples are available and it is hard to represent the negative concept uniformly. One class classifiers are more difficult to build than conventional multi-class classifier or binary classifier, for only the target classification boundary or density can be obtained when negative sample data is either absent or limited in its distribution.

One class classifiers are generally classified into three main types, which are density estimation, reconstruction and boundary estimation approaches. Two classical methods for one class classification are the density estimation method and the reconstruction method [2]. Gaussian data description, mixture of Gaussian data description and Parzen data description are well-known density estimation approaches. Some reconstruction approaches have also been developed, such as principal component analysis (PCA) data description and auto-encoded neural network data description.

Another important method for one class classification is to obtain the boundary around the target instances. Recently support vector data description (SVDD) approach has been developed to distinguish the target class from others in the pattern space [2, 5]. SVDD computes the hypersphere in the pattern space around the target class data with the minimum radius to encompass almost all the target instances and exclude the non-target ones.

There is an extensive literature on the implementation and application of SVDD. However, few researches investigate the data preprocessing method for SVDD. As we know, one class classification requires a large number of instances for object training [6]. In addition, it is difficult to decide the feature set used to find the best separation between target class and non-target class. Dimension reduction and feature selection is important for one class classification [7]. PCA preprocessing has been reported for one class classification to improve the classifier performance [8]. In fact, the feature data is often non-negative in many real life applications. Non-negative matrix factorization (NMF) is superior to PCA for non-negative data as it employs non-negative constrain which is according to the practical meaning of the real life data [9, 10]. NMF has the ability to find the hiding data structure and has been successfully used for feature extraction in some applications. In this paper a NMF and SVDD based one class classifier is developed. Experimental results demonstrate that the performance of the proposed classifier is superior to PCA based SVDD classifier and other one class classifiers.

2. NMF & SVDD Based One Class Classifier

We will introduce support vector data description firstly in this section.

Given an instance set $X = [x_{ld}] \in R^{L \times D}$, where *L* is the number of samples, and *D* is the number of features. The *i*-sample is denoted as x_i. The SVDD identifies a hypersphere with minimum volume containing all or most of the instance samples. The hypersphere volume is characterized with its center *c* and radius *R* in the new feature space. The objective of minimum volume is achieved by minimizing R^2, this constrained optimization problem can be formulated as

$$\begin{aligned} &\min \ F(c) = R^2 \\ &s.t. \ \ \|\varphi(x_i) - c\|^2 \le R^2, \qquad i = 1,...,L \end{aligned} \tag{1}$$

where $\varphi(\cdot)$ maps the feature data into a new feature space, and $\|\cdot\|$ is the L_2-norm.

To allow the possibility of outliers in the training data set, slack variables are introduced as

$$\begin{aligned} &\min \quad F(c,\xi) = R^2 + C\sum_{i=1}^{L} \xi_i \\ &s.t. \ \ \|\varphi(x_i) - c\|^2 \le R^2 + \xi_i, \ \ \xi_i \ge 0, i = 1,...,L \end{aligned} \tag{2}$$

where *C* is the penalty coefficient for outliers, and ξ_i is the distance between the *i*-th instance sample and hypersphere.

It is also possible to use a kernel $K(u,v)$ to represent the inner product. The Gaussian kernel $K(x,z) = \exp(-\|x - z\|^2 / s^2)$ is known as an efficient kernel for SVDD [5], and we always use it in this paper. The above problem can be solved by optimizing the following dual problem after introducing Lagrange multipliers

$$\begin{aligned} &\max \sum_{i=1}^{L} \alpha_i K(x_i, x_i) - \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j K(x_i, x_j) \\ &s.t. \sum_{i=1}^{L} \alpha_i = 1, \quad \alpha_i \in [0, C], i = 1,..., L \end{aligned} \tag{3}$$

A quadratic programming algorithm can be employed to solve the above problem. There are three types of training instances depending on whether $\alpha_i = 0$, $0 < \alpha_i = 0 < C$ or $\alpha_i = C$. When $\alpha_i = 0$, the instances are within the hypersphere. When $0 < \alpha_i = 0 < C$, the instances are on the hypersphere boundary. When $\alpha_i = C$, the instances are outside the hypersphere and have nonzero ξ_i. The instances are also called support vectors (SV) when $\alpha_i \ne 0$.
The hypersphere center can be obtained by

$$c = \sum_{x_i \in SV} \alpha_i \varphi(x_i) \tag{4}$$

The square radius R^2 can be calculated with the distance between *c* and any support vector *x* on the ball boundary.

$$\begin{aligned} R^2 = K(x,x) - 2\sum_{x_i \in SV} \alpha_i K(x_i, x) \\ + \sum_{x_i \in SV} \sum_{x_j \in SV} \alpha_i \alpha_j K(x_i, x_j) \end{aligned} \tag{5}$$

The above SVDD hypersphere result can be used for one class classification after training. During the classification, the sign of the following function is used to judge whether an instance is inside the SVDD hypersphere

$$\begin{aligned} D(x) &= \mathrm{sgn}(R^2 - \|\varphi(x) - c\|^2 \\ &= \mathrm{sgn}(z + 2\sum_{x_i \in SV} \alpha_i K(x_i, x) - K(x,x)) \end{aligned} \tag{6}$$

where

$$z = R^2 - \sum_{x_i \in SV} \sum_{x_j \in SV} \alpha_i \alpha_j K(x_i, x_j) \tag{7}$$

A positive sign implies that the tested instance is within the SVDD hypersphere.

Next we will give a summary to non-negative matrix factorization, and we will describe our one class classification method.

Non-negative matrix factorization was originally proposed by Paatero and Tapper [9]. Given the observation matrix

$X = [x_{ld}] \in R^{L \times D}$, and the lower-rank *J*, NMF finds such non-negative factors $A = [a_{lj}] \in R^{L \times J}$ and $T = [x_{jd}] \in R^{J \times D}$ that

$$X \cong AT \tag{8}$$

Non-negative constrains are applied to *A* and *T* during the decomposition, that means $a_{lj} \geq 0$ and $t_{jd} \geq 0$. NMF became popular after the simple multiplicative update rule provided by Lee and Seung [10]. Some other algorithms have been developed after their algorithm [11, 12]. NMF has been successfully used in a variety of real world applications, such as pattern recognition and data mining.

NMF can be solved by minimizing the difference between *X* and *AT* in terms of the squared Euclidean distance

$$f(A,T) = \|X - AT\|_F^2 / 2 \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Alternating nonnegative least squares is an efficient algorithm to solve the above problem using a block coordinate descent method in bound-constrained optimization. After initializing non-negative A^0 and T^0, the following update rule is employed

$$A^{k+1} \leftarrow \arg\min f(A^k, T^k)$$
$$T^{k+1} \leftarrow \arg\min f(A^{k+1}, T^k) \tag{10}$$

where *k*=0,1,2,.... We can use a projected gradient method called alternating non-negative least squares [1] to keep all the elements non-negative. This method shows very fast convergence and it is used in this paper.

Only non-negative data have practical physical meaning in many real world applications, and the underlying components of data with non-negative decompositions are able to provide physical interpretation. For example, the sensor measurement results of distance and volume, the housing price in the market. Non-negative decomposition provides an efficient tool to extract the relevant parts from the data. In this paper, we employ NMF in one class classifier to find the local patterns hidden in the training data, and we expect the non-negative constrains to give the natural representation of the training data with decomposed components. It is also important that NMF is an additive model which does not allow subtraction. Therefore NMF is able to describe the entire entity with the decomposed parts, that is to say NMF is a part-based representation. A zero-value represents the absence and a positive number represents the presence of the decomposed components in our one class classification application. This additive nature of NMF is expected to result in a new base of the data features.

This view of NMF leads to a two-stage one class classification method. Firstly, projected gradient method is used to perform NMF on the training data *X*. We can obtain the base *A* and the projection *T* with NMF *X*=*AT*. *A* is the new base which includes the structure and components information hidden in the training data *X*. *T* is the projection result of the training data *X* onto the base *A*. *T* can provide more feature information for one class classification. Secondly, SVDD is employed to perform one class classification training with *T* and corresponding labels. The trained SVDD can be used for one class classification after training. When we perform one class classification with test data *Q* , the test data features *P* can be obtained after the test data *Q* is projected on the base *A* with *Q*=*AP*. And the final classification results can be obtained when the trained SVDD is employed to the test data features *P*.

3. Experiments and Discussion

We will introduce support vector data description firstly in this section.

False positive rate (FPR) and false negative rate (FNR) are usually employed as error measurement for classification. FPR is the ratio of the number of non-target instances which are mistakenly classified as target to the total number of non-target instances. Similarly, FNR is the ratio of the number of target instances those are mistakenly classified as non-target to the total number of target instances. A good one class classifier will have both a small FPR and a small FNR. Recall (RC) is also widely used for classification accuracy measurement. Recall is defined as the ratio of the number of target instances those are correctly predicted to the total number of target instances. A good one class classification will have a big recall value.

Receiver-operating characteristic (ROC) curve that is a function of the true positive ratio to the false positive is usually used to compare the performance of classifiers, but the curve comparison of different classifiers is not easy. The area under the ROC curve (AUC) measure is employed to compare the performance in our experiments [13]. AUC is calculated from the ROC curve values. With this definition, the larger the AUC value is, the better is the performance of a one-class classifier.

In our experiments, data description toolbox (DDTools 1.7.5) [14] is used. And default parameters for this toolbox are employed. The tolerance for NMF is 10, and the maximum iteration number is 5. 70 percent of the whole instances is selected as the training data, and the other instances data is used to evaluate the classification performance after the training.

3.1 Compare with Other One Class Classifiers

Footnotes should be typed in singled-line spacing at the bottom of the page and column where it is cited. Footnotes should be rare.

Databases with non-negative data used widely for classification test in other literature [16, 17] were employed to test the performance of the proposed algorithm in this paper. Results on the wine recognition database and the Boston housing database from University of California Irvine machine learning repository [15] are reported here. The wine recognition database is used to determine the origin of wines using chemical analysis. The class 1 is regarded as target class in this wine recognition database. There are 59 target instances and 119 outlier instances with 13 features in the wine recognition database. The Boston Housing database is used to predict the housing price in suburbs of Boston. The class whose median price is less than 35000 dollar is regarded as target class. There are 458 target instances and 48 outlier instances with 13 features in the Boston Housing database. 70 percent target instances and outlier instances are randomly selected from the whole instances data for classifier training, and the other 30 percent for testing and evaluation in our experiments.

Different one class classifiers were evaluated on the wine database and the housing database. These classifiers included Gauss, mixture Gauss (MixGauss), PCA, SVDD, Parzen, auto-encoded neural network (AENN), and the proposed one class classifier. Experimental results of the wine database and the housing database are listed in Table 1 and Table 2, respectively. The factor size of *J* in equation (8) of NMF for the wine database and the housing database is selected as 6 and 5, respectively. Both SVDD method and Parzen method can accurately classify all the target instances for the wine database in Table 1, but they all have bad FPR value. This means these two methods classified most outlier instances as target. The wrong result is caused by the insufficient information in the training set to correctly estimate the parameters for the classifier. The error classification rates for target and outlier of Gauss, mixture Gauss, PCA, and auto-encoded neural network are higher than 11 percent. All the error classification rates of the proposed method are less than 6 percent. The proposed method obtained the less total error classification rate and the greater recall value. The AUC value gives the overall performance evaluation. The proposed method obtained the greatest AUC value, it has the superior performance. Similar results can also be found in Table 2 for the housing database. The proposed method obtained the best performance compared with other methods in our experiments.

Table 1: Comparison of one class classifiers for wine database

Method	*FPR*	*FNR*	*RC*	*AUC*
Gauss	0.0292	1.0000	0.9708	0.6241
MixGauss	0.0511	1.0000	0.9489	0.6554
PCA	0.0438	0.9286	0.9562	0.6481
SVDD	0.9781	0.0000	1.0000	0.5615
Parzen	0.8248	0.0000	0.1752	0.5865
AENN	0.0438	0.7857	0.9562	0.5553
Proposed	0.2190	0.4286	0.7810	0.8186

Table 2: Comparison of one class classifiers for housing database

Method	*FPR*	*FNR*	*RC*	*AUC*
Gauss	0.1765	0.2286	0.8235	0.9109
MixGauss	0.2941	0.1714	0.7059	0.8588
PCA	0.1765	0.7429	0.8235	0.6084
SVDD	1.0000	0.0000	0.0000	0.6151
Parzen	0.7647	0.0000	0.2353	0.7647
AENN	0.1176	0.3143	0.8842	0.8336
Proposed	0.0588	0.0286	0.9412	0.9950

3.2 Compare with PCA Based SVDD Classifier

PCA is often used for feature extraction before classification. It can capture the data variance in the squared error sense and map data into orthonormal subspace. Eigenvalue decomposition is used to obtain the eigenvectors of the target covariance matrix in the practical calculation. The eigenvectors corresponding to the largest eigenvalues are considered as the principal components, they are the principal axis in the direction of the largest variance. These eigenvectors are used to form an orthonormal basis for data mapping. The number of the orthonormal basis vectors is optimized to represent the data variance.

The one class classifiers that combine PCA feature extraction and SVDD are also compared with the proposed method in our experiments. Classical PCA and kernel PCA (KPCA) [8] based feature selection methods are used for test. The feature number was automatically optimized in the experiment while the total variance was selected as 90 percent.

Experimental results of the wine database and the housing database are listed in Table 3 and Table 4, respectively. The overall performance evaluation provided by AUC value shows that PCA and kernel PCA based SVDD classifiers can improve the classification performance. However, the performance improvements of these methods are small compared with the proposed method.

Databases with non-negative data used widely for classification test in other literature [16, 17] were employed to test the performance of the proposed algorithm in this paper. Results on the wine recognition database and the Boston housing database from University of California Irvine machine learning repository [15] are reported here. The wine recognition database is used to determine the origin of wines using chemical analysis. The class 1 is regarded as target class in this wine recognition database. There are 59 target instances and 119 outlier instances with 13 features in the wine recognition database.

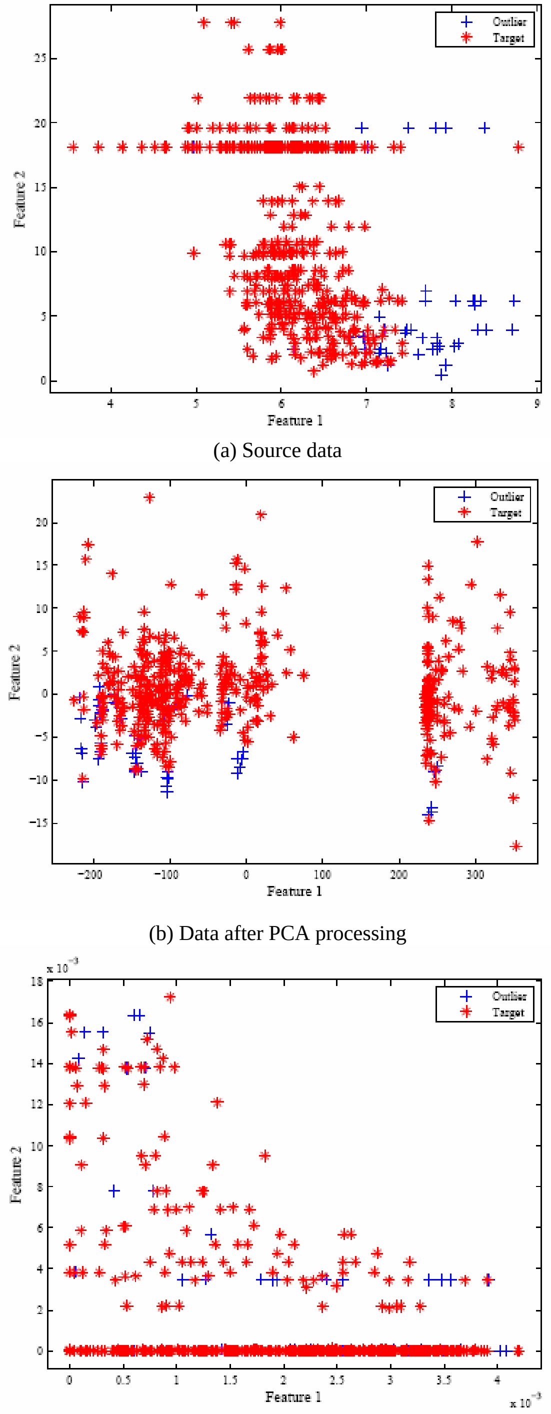
Table 3: Compare with PCA based SVDD for wine database

Method	***FPR***	***FNR***	***RC***	***AUC***
SVDD	1.0000	0.0000	0.0000	0.6151
PCA+SVDD	0.9412	0.0286	0.0588	0.6630
KPCA+SVDD	1.0000	0.0000	0.0000	0.6206
Proposed	0.0588	0.0286	0.9412	0.9950

Table 4: Compare with PCA based SVDD for housing database

Method	***FPR***	***FNR***	***RC***	***AUC***
SVDD	0.9781	0.0000	0.0219	0.5615
PCA+SVDD	0.5547	0.3571	0.4453	0.5928
KPCA+SVDD	1.0000	0.0000	0.0000	0.6019
Proposed	0.2190	0.4286	0.7810	0.8186

The data feature distribution of the housing database is plotted in Fig.1. The two principal features those are obtained with PCA analysis method of the source data, PCA processed data and NMF processed data is plotted in 1(a), 1(b) and 1(c), respectively. Some outlier instances of source data are in the high density area of the target instances as showed in Fig.1(a). This is similar for PCA processed data as showed in Fig.1(b). And most outlier instances are distributed in the low density area of the target instances area for NMF processed data as showed in



(a) Source data

(b) Data after PCA processing

(c) Data after NMF processing

Fig. 1 Data feature plot of housing database

Fig.1(c). This showed that the new feature space obtained with NMF based processing method is more suitable for one class classification. PCA decomposition satisfies with orthogonal constrain, and cannot assure the decomposition results are non-negative. NMF satisfies with the non-negative constrain, and it is more suitable for non-negative data application to find the hidden structure. This is the reason that NMF is superior to PCA in one class classification for non-negative data.

3.3 Comparison of different factorization sizes

Deciding the factorized matrix size *J* in equation (8) is important for NMF. It is known that when the size of *X* is L×D, *J* needs to satisfy $J \leq (L \times D)/(L+D)$. Different *J* in the proposed method is employed to the wine database and the housing database, and the results are listed in Table 5 and Table 6, respectively. Compared with the AUC value 0.6151 of SVDD method in Table 1 for the wine database, all the AUC value of different *J* in Table 5 are greater. This means our proposed NMF based method is efficient to improve the performance of SVDD for one class classification. This can also be verified in Table 6 for the housing database, where all the AUC value of different *J* is also greater than original SVDD method. But the best choice of *J* is application dependent, parameter optimization can be employed for selection of *J*.

Table 5: Comparison of different factorized size for wine database

Size of J	*FPR*	*FNR*	*RC*	*AUC*
J=5	0.1176	0.1429	0.8824	0.9664
J=6	0.0588	0.0288	0.9412	0.9950
J=7	0.1765	0.1429	0.8235	0.9529
J=8	0.1765	0.1429	0.8235	0.9513
J=9	0.0588	0.4000	0.9412	0.6840
J=10	0.1765	0.5714	0.8235	0.6975

Table 6: Comparison of different factorized size for housing database

Size of J	*FPR*	*FNR*	*RC*	*AUC*
J=5	0.2190	0.4286	0.7810	0.8186
J=6	0.3066	0.3571	0.6934	0.7492
J=7	0.2044	0.4286	0.7956	0.8175
J=8	0.1606	0.6429	0.8394	0.7044
J=9	0.2628	0.3571	0.7372	0.8149
J=10	0.1898	0.2857	0.8102	0.7753

4. Conclusions

A two stage method for one class classification employing NMF and SVDD for non-negative data is proposed. NMF is used to project sample instances to a new feature space before SVDD is employed for classification. There are several advantages in this hybrid method. First, NMF is more efficient than PCA to find the hidden structure for non-negative data, and the feature space produced with NMF is appropriate for SVDD. Second, the proposed method is superior to other classical one class classifiers.

Acknowledgments

This work is partially supported by China Postdoctoral Science Foundation (20100471057), Shandong Provincial Natural Science Foundation (ZR2011FM005), Promotive Research Fund for Excellent Young and Middle-Aged Scientists of Shandong Province (BS2010DX001), Science and Technique Development Project of Weihai (2010-3-96) and Key Lab Open Fund of Harbin Institute of Technology (HIT.KLOF.2012.078).

References

[1] M. Moya, M. Koch and L. Hostetler, "One-class classifier network for target recognition applications", in Proc. World Congress on Neural Networks, 1993, pp.791-801.

[2] D.M.J. Tax, "One class classification", Ph.D. thesis, Delft University of Technology, Delft, Netherlands, 2001.

[3] P. Juszczak, D.M.J. Tax, E. Pekalska and R.P.W. Duin, "Minimum spanning tree based one-class classifier ", Neurocomputing, 2009, Vol. 72, pp. 1859-1869.

[4] L. Jamali, M. Bazmara, and S. Jarari, "Feature selection in imbalanced data sets", International Journal of Computer Science Issues, 2012, Vol. 9, No. 3, pp. 42-45.

[5] D.M.J. Tax and R.P.W. DuinP. "Support vector domain description", Pattern Recognition Letters, 1999, Vol. 20, No. 11-12, pp. 1191-1199.

[6] H. Yu, "Single-class classification with mapping convergence", Machine Learning, 2005, Vol. 61, No.1-3, pp. 49-69.

[7] S. D. Villalba and P. Cunningham, "An evaluation of dimension reduction techniques for one-class classfication", Artificial Intelligence Review, 2007, Vol. 27, No. 4, pp. 273-294.

[8] D.M.J. Tax and P. JuszczakA. "Kernel whitening for one-class classification", International Journal of Pattern Recognition and Artificial Intelligence, 2003, Vol. 17, pp. 333-347.

[9] U. Paatero and A. Tapper, "Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values", Envirometrics, 1994, Vol. 5, No. 2, pp. 111-126.

[10] D.D. Lee and H.S. Seung, "Learning the parts of objects by nonnegative matrix factorization", Nature, 1999, Vol.401, pp.788-791.

[11] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization", Neural Computation, 2007, Vol.19, No.10, pp.2756-2779.

[12] A. Cichocki, R. Zdunek, A. Phan and S. Amari, Nonnegative matrix and tensor factorizations, John Wiley, Singapore, 2009.

[13] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning apgorithms", Pattern Recognition, 1997, Vol.30, No.7, pp.1145-1159.

[14] D.M.J. Tax, DDTools, The data description toolbox for Matlab, http://prlab.tudelft.nl/david-tax/dd tools.html, 2010.

[15] A. Frank and A. Asuncion, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, University of California, 2010.

[16] M. Breaban and H. Luchian, "A unifying criterion for unsupervised clustering and feature selection, Pattern Recognition", 2011, Vol.44, No.4, pp.854-865.

[17] M. Kalakech, P. Biela, L. Macaire and D. Hamad, "Constraint scores for semi-supervised feature selection: A comparative study", Pattern Recognition Letters, 2011, Vol.32, No.5, pp.656-665.

Liyong Ma . received the B.Sc degree from Harbin Institute of Technology, Harbin, China, in 1993, the M.Sc. degree from Harbin University of Science and Technology, Harbin, China, in 1996, and Ph.D degree from Harbin Institute of Technology in 2007, respectively. He is currently an Associate Professor at Harbin Institute of Technology at Weihai, Weihai, China. His main research areas include intelligent testing and information processing, biomedical imaging and image processing.

Naizhang Feng received the B.Sc. and Ph.D. degree in control engineering from Harbin Institute of Technology, Harbin, China, in 1998 and 2005, respectively. From 2005 to 2007, he was a post-doctor in Fudan University, Shanghai, China. He is currently a Professor in Harbin Institute of Technology at Weihai. His research interests include signal detection and information processing and medical ultrasound imaging.

Qi Wang received the B.Sc. and M.Sc. degree in electromagnetic measurement from Harbin Institute of Technology, Harbin, China, in 1967 and 1980, respectively. From 1985 to 1987 and 1993 to 1995, he visited the University of Tsukuba and the Chiba Institute of Technology, Japan, as a Researcher. He is currently a Professor at the Harbin Institute of Technology. So far, he has more than 120 papers published. His research interests include the intelligent testing instrumentation, information processing, sensor fault diagnosis, and sensor fusion.