A Novel Method of Feature Selection based on SVM

Quanjin Liu
College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing, China
School of Physics & Electronic Engineering, Anqing Normal College, Anqing, China
Email: liuquanjing666@126.com

Zhimin Zhao
College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing, China
Email: nuaazhzhm@126.com

Ying-xin Li
Institute of Machine Vision and Machine Intelligence, Beijing Jingwei Textile Machinery New Technology Co., Ltd., Beijing, China
E-mail: linterlee@gmail.com

Xiaolei Yu
Jiangsu Institute of Standardization, Nanjing, China
E-mail: nuaaxiaoleiyu@126.com

Yong Wang
The Second Affiliated Hospital, Anhui Medical University, China
E-mail: yongwangefy@sina.com

***Abstract*—A novel method of feature selection combined with sample selection is proposed to select discriminant features in this paper. Based on support vector machine trained on training set, the samples excluding the misclassified samples and support vector samples are used to select informative features during the procedure of recursive feature selection. The feature selection method is applied to seven datasets, and the classification results of the selected discriminant features show that the method is effective and reliable for selecting features with high classification information.**

***Index Terms*—Discriminant feature selection, Support Vector Machine, Sample selection, Gene expression profile**

I. Introduction

Feature selection method is one of the techniques to reduce feature dimension for classification in machine learning[1,2,3,4]. The method can be categorized into the filter method, the wrapper method and the embedded method. The former method is independent of classifier and select key features by the divisibility index of samples[5]. In other words, the wrapper method selects informative features based on classifier. And the embedded method selects informative features in the learning time. These methods are often integrated to extract the discriminant features for classification from high-dimensional datasets [6].

Many literatures focused on selecting key genes from gene expression profile dataset [7]. In [8], Guyon et al. selected critical genes in the process of recursive feature elimination based on the feature selection method (RFE-SVM). On the other hand, clustering algorithm was also used to select critical genes from high dimensional dataset[9]. In [10], Liu et al. proposed a feature selection method based on the fuzzy clustering algorithm (FS-CLUSTER) [10].

Information of the support vector samples (SVs) in SVM model is used for construct the classification decision function. Literature [11] thinked it is inappropriate to carry out classification based on SVs on the dataset with the uneven number of heterogeneous samples. Lyhyaoui et al. selected two samples with the nearest distance between each cluster to establish the classifier[12].

This paper proposes a Sample Selection Method (SSM) to select discriminant features form high dimensional dataset. We select informative features from 7 datasets based on FS-CLUSTER and SSM (FS-SSM). The original dataset is divided into the training set, validation set and independent test set randomly. FS-SSM is conducted on the samples selected by SSM from the training set. Experimental results on the 7 datasets demonstrate that SSM method is useful to improve the performance of FS-SSM based on clustering model.

The rest of this paper is organized as follows. Section 2 proposes the sample selection method (SSM). Section 3 introduces the feature selection method (FS-SSM).

doi:10.4304/jcp.8.8.2144-2149

Section 4 describes the feature selection experiments on the 7 datasets. Section 5 concludes the paper.

II. Sample Selection Method (SSM)

Vapnik proposed support vector machine (SVM) algorithm based on the statistical learning theory and the structural risk minimization principle[13]. SVM is a machine learning algorithm which can get good generalization ability in the case of dataset with limited samples [14,15].

Let $X_i = \{x_{i1}, ..., x_{ij}, ..., x_{im}\}$ be a sample of training set X and $y_i \in \{+1, -1\}$ be a class label of X_i. If the samples can be divided into 2 groups by the the hyperplane $g(x) = \omega \cdot x + b = 0$, $g(x) = \omega \cdot x + b$ is regarded as the linear discriminant function. The margin between the pair of parallel hyperplanes $\omega \cdot x + b = \pm 1$ is determined by ω. Quadratic programming algorithm is used to seek the maximum margin:

$$\text{minimize: } \Phi(\omega) = \frac{1}{2}\|\omega\|^2 \qquad (1)$$
$$\text{subject to: } y_i(\omega^T \cdot X_i + b) \geq 1, i = 1, 2, ..., n$$

This quadratic programming problem can be solved by Lagrange multipliers algorithm. And the discriminant function is obtained as follows:

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{sv} \lambda_i^* y_i (X_i \cdot x) + b^*\right\} \qquad (2)$$

The sample X_i, which lies on the parallel hyperplanes, is the support vector sample (SV) [16]. λ_i^* is Lagrange multipliers of SV and sv is the number of support vector samples. The SVs on the parallel hyperplanes can be considered as the border of each class.

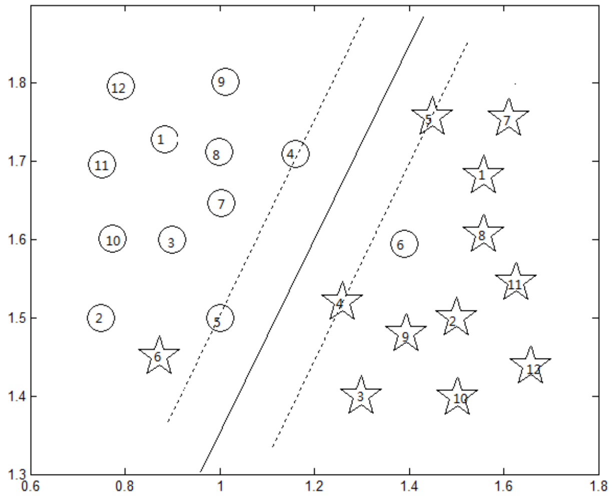

Figure 1. Illustration of two types of samples.

As shown in Figure 1, 24 samples marked by the circular and pentagram icons belong to the negative and positive class, respectively. The line in the middle refers to the optimal discriminant line and the dashed lines are 'support lines' of negative and positive class respectively. 11 samples of the 12 negative ones are at the upper left corner, and 11 samples of the 12 positive ones are at lower right corner. The NO. 4 and 5 samples marked by circle icons on the upper support line are negative SVs. The NO. 4 and 5 samples marked by pentagram icons on the lower support line are positive SVs. The negative NO. 6 sample and positive NO. 6 sample on the wrong side are misclassified by the optimal discriminant line.

The figure demonstrates that if the misclassified samples (MSs) and SVs are removed, the margin between the heterogeneous samples can be increased. On condition that the MSs and SVs only account for a small part of the training set, the removal of the samples will not change the original information structure of dataset. So, if MSs and SVs are deleted from training set, not only the margin between different classes can be expanded but also the class information of original dataset can be retained. And the distance among samples of the same class can also be shortened relatively.

The compactness within-class and dispersion between classes are important index for feature selection [1].We proposed sample selection method (SSM), which selects samples other than MSs and SVs after training SVM, for feature selection.

III. Feature Selection Method Based on SSM (FS-SSM)

Fuzzy Interactive Self-Organizing Data Algorithm (ISODATA) is a kind of clustering algorithm with simple structure and high running speed [17,18].

The samples in training set belongs to ***2*** clusters and membership u_{ki} of sample X_i implies the relationship between features of X_i and k^{th} class [18,19]. Literature [10] defines sensitivity formula of the j^{th} feature of samples to the membership:

$$D(j) = \sum_{k=1}^{2} |D(k,j)| = \sum_{k=1}^{2} \left| \sum_{i=1}^{n} \sum_{p=1}^{n} \frac{\partial u_{ki}}{\partial x_{pj}} \right| \qquad (3)$$

$D(k, j)$ reflects the contribution of the j^{th} feature to the k^{th} cluster and $D(j)$ can be considered as the key factor for fuzzy ISODATA Clustering.

FS-CLUSTER method selects informative features based on the sensitivity factor defined as Eq (3). The "cluster" in fuzzy ISODATA algorithm implies the underlying structure of the dataset. The discriminant function constructed by the selected features has the high recognition ability.

Removing MSs and SVs can increase margin between different classes of samples relatively and the other samples would have higher membership in clustering algorithm. We propose a novel feature selection method based on FS-CLUSTER and SSM (FS-SSM).

FS-SSM algorithm is shown in Figure 2. During the process of recursive feature selection, FS-SSM generates candidate feature subsets based on the samples selected by SSM in the training set. Classification performance of the candidate subsets is evaluated by SVM [20] and K nearest neighbor (KNN) [21] classifiers, which are trained on training set. Class information of the candidate subsets is quantified by AUC value (Area Under the receiver operating characteristic Curve) [22,23] and the correct recognition rate of classification on validation set. The subset with the best classification performance is regarded as the optimal feature subset with the most class information.

```
FS-SSM algorithm:
Step1:Train SVM model based on the feature subset $F_{subset}$ on
      training set;
Step2:Fuzzy ISODATA cluster and generate new $F_{subset}$ based
      on the samples selected by SSM;
Step3:Evaluate classification performance of $F_{subset}$ by SVM
      and KNN on validation set.
Step4:If dimension( $F_{subset}$ ) >3, Goto Step 1;
Step5:Select the $F_{subset}$ with the best classification performance
      as the optimal feature subset;
Step6:End.
```

Figure 2. The flowchart of FS-SSM algorithm.

IV. Feature Selection Experiment

The FS-SSM method proposed in this paper is applied to 7 datasets. As shown in Figure 2, informative features are selected based on samples selected by SSM during recursive feature selection process. To evaluate the impact of SSM method on FS-SSM, FS-CLUSETR is also carried out on the same datasets.

A. Datasets

7 datasets, whose samples belong to 2 classes, are used in feature selection experiments. As shown in Table I, Ionosphere and Promoters datasets are downloaded from Machine learning repository of University of California Irvine. The other five datasets are gene expression profile datasets.

TABLE I.
Descriptions of dataset in experiments

No.	Dataset	Features	Instances	Select scope	Reference
1	Ionosphere	34	351	34	[24]
2	Promoters	57	106	57	[24]
3	Multiple myeloma	7129	105	100	[25]
4	Acute Leukemia	7129	72	100	[7]
5	Colon	2000	62	500	[9]
6	DLBCL	7129	77	1000	[26]
7	Prostate	12600	102	1000	[27]

The second column in Table I illustrates dataset name, the third column lists the number of features, and the fourth column shows the number of samples.

The datasets are randomly divided into three parts, training set, validation set and independent test set based on the proportion of 3:1:1 in feature selection experiments.

FS-SSM and FS-CLUSTER are carried on the same training set and validation set to select optimal feature subsets. Then, the selected subsets are evaluated on the same independent test set.

B. Irrelevant Features Filtering

There are many noise and irrelevant genes in the five gene expression profile datasets with high dimensional features [1, 10]. Bhattacharyya distance [1,5] between the two types of samples is used as criteria to filter the noise and irrelevant genes before the feature selection process. The fifth column in Table I lists scope of further feature selection.

C. Optimal Feature Subset Slection

The feature selection experiments are conducted with MATLAB on a PC with 3.2 GHz Intel Core i5-3470 CPU and 4.0 GB RAM.

We set r = 2, s = 2, and ε = 0.0001 for fuzzy ISODATA algorithm in feature selection experiments. The kernel function of SVM is set as linear function and number of neighbors is set as 5 in KNN algorithm.

FS-SSM and FS-CLUSTER algorithms generate the nested candidate feature subsets on the training sets respectively. The candidate subset with the highest AUC value and recognition rate on validation sets is treated as the optimal feature subset.

To compare performance of FS-SSM and FS-CLUSTER, we conduct the two methods 40 times on each datasets. That is, each dataset is randomly divided 40 times and 40 optimal feature subsets are selected during 40 rounds of feature selection process based on the different training sets and validation sets.

D. Optimal Feature Subsets Comparision

Features in optimal feature subset are regarded as discriminant features for classification. The more class information the optimal feature subset has, the power classification ability it has. Therefore, the better the classification result of the optimal feature subset is, the higher the feature selection method performance will be.

Based on the optimal feature subset, SVM and KNN classifiers trained in the training set are used to classify the samples in the independent test set. The samples in the independent test set are independent of the validation set. The higher the AUC value and the recognition rate are, the higher the classification performance of the optimal feature subset will be.

TABLE II.
Performance of the optimal feature subsets on SVM classifier

Dataset	FS-SSM		FS-CLUSTER	
	AUC	Recognition rate	AUC	Recognition rate
Ionosphere	0.910±0.044	0.886±0.038	0.898±0.039	0.859±0.045
Promoters	0.848±0.090	0.767±0.113	0.808±0.085	0.731±0.086

Multiple Myeloma	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Acute Leukemia	0.966±0.052	0.946±0.061	0.943±0.062	0.914±0.082
Colon	0.846±0.133	0.736±0.090	0.825±0.109	0.706±0.105
DLBCL	0.891±0.115	0.861±0.109	0.875±0.080	0.856±0.104
Prostate	0.881±0.089	0.808±0.099	0.882±0.090	0.810±0.099

TABLE III.
PERFORMANCE OF THE OPTIMAL FEATURE SUBSETS ON KNN CLASSIFIER

Dataset	FS-SSM		FS-CLUSTER	
	AUC	Recognition rate	AUC	Recognition rate
Ionosphere	0.850±0.063	0.859±0.057	0.855±0.063	0.845±0.053
Promoters	0.853±0.091	0.757±0.101	0.812±0.088	0.721±0.096
Multiple Myeloma	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Acute Leukemia	0.941±0.070	0.923±0.082	0.931±0.067	0.914±0.072
Colon	0.898±0.090	0.845±0.079	0.866±0.106	0.810±0.089
DLBCL	0.888±0.094	0.859±0.103	0.873±0.064	0.829±0.087
Prostate	0.890±0.065	0.933±0.158	0.876±0.072	0.918±0.183

Performance of FS-SSM and FS-CLUSTER are compared in terms of AUC value and recognition rate of the selected optimal feature subsets in the independent tests.

Table II and Table III list the classification results of SVM and KNN, respectively. The first column is the name of datasets. The second and third columns list the mean and standard deviation of AUC value and recognition rate of the 40 optimal subsets selected by FS-SSM. Similarly, the fourth and fifth columns list classification results of the optimal subsets selected by FS-CLUSTER.

From Table II and Table III, we can know all optimal feature subsets selected by FS-SSM and FS-CLUSTER from Multiple Myeloma dataset can correctly classify all samples of the independent test set.

Figure 3 and Figure 4 illustrate the classification performance of the optimal feature subsets selected by the 2 methods form 7 datasets during the 40 rounds of feature selection processes. The x-axis indicates the datasets and the y-axis presents classification performance of the optimal subsets selected by FS-SSM and FS-CLUSTER. Height of the vertical bars presents the mean of AUC and recognition rate of the selected informative subsets.

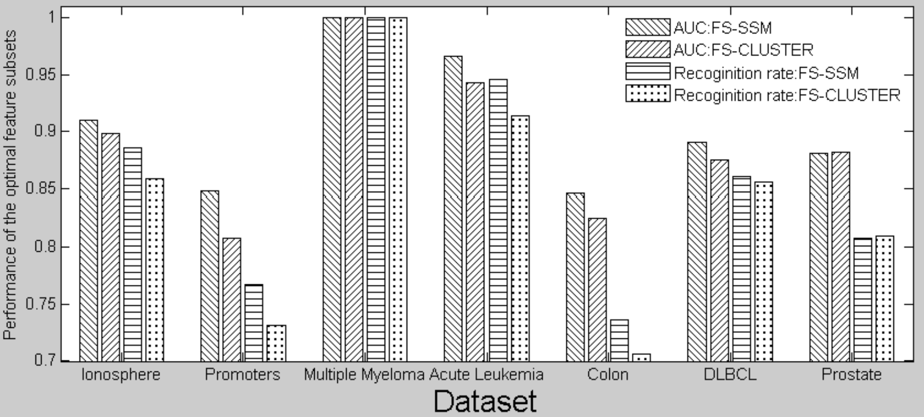

Figure 3. Performance comparision of the optimal feature subsets on SVM classifier

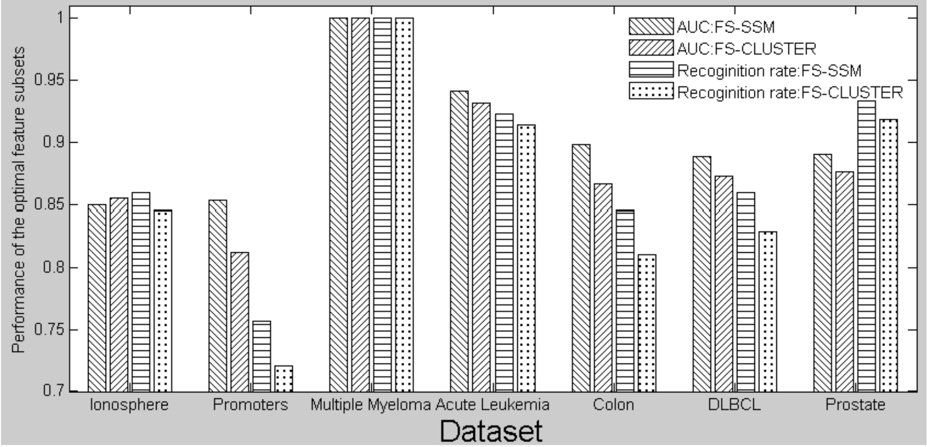

Figure 4. Performance comparision of the optimal feature subsets on KNN classifier

As shown in Figure 3, classification performance on SVM classifier of the informative feature subsets selected by FS-SSM is higher than that selected by FS-CLUSTER, except that on Prostate dataset.

From Figure 4, we can find that AUC value on KNN classifier of the optimal feature subsets selected by FS-CLUSTER is higher than that selected by FS-SSM. The other bars illustrate that KNN classification result of the optimal feature subsets selected by FS-SSM is better than that selected by FS-CLUSTER.

The independent test results show the classification performance of the informative feature subsets selected by FS-SSM is superior to that selected by FS-CLUSTER. It demonstrates that the FS-SSM method can select critical features with more class information.

The feature selection experiments prove that SSM algorithm can improve the performance of FS-CLUSTER on the 5 gene expression profile datasets. It means that SSM can be applied to key gene selection for cancer diagnosis.

As seen from the results of 40 rounds of feature selection experiments, the removal of MSs and SVs can expand the margin between heterogeneous samples, enhance the cohesion of within-class samples and improve the classification performance of the selected critical features.

V. Conclusions

This paper proposes a method of feature selection based on sample selection method (FS-SSM). Feature selection experiments on 7 datasets demonstrate the optimal feature subsets selected by FS-SSM achieve high classification performance in independent tests. Results imply the SSM is able to improve the performance of FS-SSM effectively and prove the proposed FS-SSM method has potential application on selecting critical genes for tumor diagnosis.

Experimental results also prove that removing misclassified samples and samples on class border can improve the performance of feature selection method based on clustering algorithm.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NO.10172043, NO.61173068), the Specialized Research Fund for the Doctoral Program of Higher Education of China (NO.20093218110024), International Science and Technology Cooperation Grant (NO.BZ2010060), Key Program of the Jiangsu Municipal Bureau of Quality and Technical Supervision (NO. KJ122714) and the Natural Science key Foundation of the Department of Education of Anhui Province of China (NO.KJ2010A226).

References

[1] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Academic Press, New York, 1999.

[2] D. P. Zhang and J. Deng, “The Data Mining of the Human Resources Data Warehouse in University Based on Association Rule,” Journal of computers, vol. 6(1), 2011, pp.139-147.

[3] T. Dong, W. Q. Shang and H. B. Zhu, “An Improved Algorithm of Bayesian Text Categorization,” Journal of Software, vol. 6(9), 2011, pp. 1837-1843.

[4] W. Yang, K. Q. Wang and W. M. Zuo, “Neighborhood Component Feature Selection for High-Dimensional Data,” Journal of computers, vol. 7(1), 2012, pp. 161-168.

[5] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification. 2nd, John Wiley & Sons, New York, 2001.

[6] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” Journal of Machine Learning Research, vol. 3, 2003, pp.1157–1182.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek and J. P. Mesirov, et al., “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,” Science, vol. 286(5439), 1999, pp.531-537.

[8] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46(13),2000, pp.389-242.

[9] U. Alon, N. Barkai, DA. Notterman, K. Gish, S. Ybarra and D. Mack, et al., “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” PNAS USA, vol. 96(12), pp.6745-6750, 1999.

[10] Q. J. Liu, Z. M. Zhao, Li, Y-X. and Y. Y. Li, ,“Feature Selection Based on Sensitivity Analysis of Fuzzy ISODATA,” Neurocomputing, vol. 85, 2012, pp.29-37.

[11] R. Akbani, S. Kwek and N. Japkowicz, “Applying support vector machines to imbalanced datasets,” In: J.-F. Boulicaut et al. (Eds.): ECML 2004, LNAI 3201, 2004, pp.39-50.

[12] A. Lyhyaoui, M. Martínez, I. Mora, M. Vázquez, J. Sancho and A. R. Figueiras-Vidal, “Sample selection via clustering to construct support vector-like classifiers,” IEEE Trans. Neural Networks, vol. 10(6), 1999, pp.1474-1481.

[13] V. N. Vapnic, Statistical Learning Theory. John Wiley and Sons, New York, 1998.

[14] Y. Wei and X. Wu, “A New Fuzzy SVM based on the Posterior Probability Weighting Membership,” Journal of computers, vol. 7(6), 2012, pp. 1385-1392.

[15] S. X. Yang and G. y. Yang, “Emotion Recognition of EMG Based on Improved L-M BP Neural Network and SVM,” Journal of Software, vol. 6(8), 2011, pp. 1529-1536.

[16] V. N. Vapnic, The Nature of Statistical Learning Theory. Springer, New York, 1996.

[17] J. C. Bezdek, “Physical interpretation of fuzzy ISODATA. EEE SMC,” SMC-6, 1976, pp.387-390.

[18] F. Marcelloni, “Feature selection based on a modified fuzzy C-means algorithm with supervision,” Information Sciences, vol. 151, 2003, pp.201-226.

[19] J. C.Bezdek , Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[20] J. P. Zhang, J. F. F. He, L. Ma and J. Y. Y. Li, “ORPSW: a new classifier for gene expression ata based on optimal risk and preventive patterns,” Journal of computers, vol. 6(6), 2011, pp. 1198-1205.

[21] H.-H. Hsu, A. C. Yang and M.-D. Lu ,"KNN-DTW Based Missing Value Imputation for Microarray Time Series Data," Journal of computers, vol. 6(3), 2011, pp. 418-425.

[22] Y-X Li,, S. Ji, S. Kumar, J. Ye and Zhou, Z-H, "Drosophila gene expression pattern annotation through multi-instance multi-label learning," ACM/IEEE Transactions on Computational Biology and Bioinformatics, vol. 9(1), 2012, pp.98-112.

[23] F. Provost and T. Fawcett, "Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions," In: Proc. Third Internat. Conf. on Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA, pp.43–48, 1997.

[24] UC Irvine. Machine Learning Repository [DB/OL]. http://archive.ics.uci.edu/ml/datasets.html.

[25] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok and R. C. Aguiar et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," Nat. Med. , vol. 8(1), 2002, pp.68-74.

[26] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola and C. Ladd et al., "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol. 1(2), 2002, pp.203-209.

[27] F. Zhan, J. Hardin, B. Kordsmeier, K. Bumm, M. Z. Zheng and E. Tian et al., "Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells," Blood, vol. 99(5), 2002, pp.1745-1757.

Quanjin Liu is a Ph.D. candidate in College of Science, Nanjing University of Aeronautics and Astronautics, China. His research interests include machine learning, data mining and bioinformatics.

Zhimin Zhao received the MS degree from NUAA in 1992. She was selected as the Fellow of Chinese Optical Society in 2008. She is now a Full Professor in College of Science, NUAA. Her research interests include advanced measurement, control, and intelligent computation.

Ying-Xin Li received the PhD degree in pattern recognition and intelligent systems from Beijing University of Technology, China, in 2006. His research interests include machine learning, data mining, bioinformatics, and machine vision.

Xiaolei Yu received the PhD degree in 2011 from NUAA. His research interest is in the area of automatic control, information technology, automatic identification and information integration.

Yong Wang received the M.D. degree in Shuzhou University, China, in 2006. His research interest is clinical treatment of cancer.