academicJournals

Vol. 8(17), pp. 809-816, 9 May, 2013
DOI: 10.5897/IJPS12.327
ISSN 1992-1950 © 2013 Academic Journals
http://www.academicjournals.org/IJPS

International Journal of Physical Sciences

Full Length Research Paper

A method of dual-process sample selection for feature selection on gene expression data

Quanjin Liu[1,2], Zhimin Zhao[1]*, Ying-xin Li[3] and Xiaolei Yu[4]

[1]College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China.
[2]School of Physics and Electronic Engineering, Anqing Normal College, Anqing 246011, China.
[3]Institute of Machine Vision and Machine Intelligence, Beijing Jingwei Textile Machinery New Technology Co., Ltd., Beijing 100176, China.
[4]Jiangsu Institute of Standardization, Nanjing 210029, China.

Accepted 23 August, 2012

A method of dual-process sample selection based on support vector machine (SVM) is proposed to select informative features in this paper. Samples in a training set are used to train a SVM model, and the samples excluding support vector samples are chosen to select critical features in the procedure of recursive feature elimination (RFE). The effect of the dual-process sample selection method on feature selection is evaluated using the classification and the clustering performance of the selected features. The proposed dual-process sample selection method is applied to five gene expression datasets, and the experimental results show that the method is useful to improve the performance of the feature selection method based on fuzzy interactive self-organizing data algorithm (ISODATA). This indicates the method is reliable and effective for selecting informative genes from gene expression data.

Key words: Feature selection, support vector machine, fuzzy interactive self-organizing data algorithm (ISODATA), dual-process sample selection.

INTRODUCTION

Feature selection method removes irrelevant features and selects a small portion of features with strong classification ability from the original dataset (Theodoridis and Koutroumba, 1999). Depending on the classification models used, feature selection method can be classified into the “filter” and the “wrapper” method. The former method takes the characteristics of the dataset itself into account and utilizes the divisibility index of samples to select informative features (Duda et al., 2001; Uncua and Türkşenb, 2007). In contrast, the latter method conducts the feature selection based on some classification models. The integration of those two methods is the major focus of the research on the feature selection from high-dimensional datasets (Guyon and Elisseeff, 2003; Jensen and Shen, 2009).

For cancer classification and diagnosis, many literatures studied how to select informative genes from microarray dataset, which has thousands of genes and only dozens of samples. On one hand, filter method was used to filtered the "irrelevant" genes and select critical genes (Golub et al., 1999; Lan and Vucetic, 2011). On the other hand, wrapper methods were applied to informative genes selection. For instance, Guyon et al. (2000) proposed the feature selection method (SVM-RFE) based on SVM to select critical genes in the process of recursive feature elimination. Tang et al. (2008) designed a recursive fuzzy granular support vector machine to select informative genes for cancer diagnosis. In addition,

*Corresponding author. E-mail: nuaazhzhm@126.com. Tel: 86-25-84892011. Fax: 86-25-84892011.

ensemble method was applied to wrapper method based on the difference between gene subsets (Abeel et al., 2010). Meanwhile, unsupervised learning algorithm was also used to analyze microarray dataset and select discriminant genes (Alon et al., 1999). Liu et al. (2012) proposed the feature selection method based on fuzzy Interactive Self-Organizing Data Algorithm (RFE-ISODATA) and selected informative genes from 5 cancer microarray datasets.

Sample selection method is to select the key sample to build the classification decision function. SVM only uses the information of the support vector samples (SVs) for the classification decision, thus it can work with a high speed. However, for the dataset with the uneven number of heterogeneous samples, it is inappropriate to use SVs to carry out classification (Akbani et al., 2004). Tang et al. (2009) later improved the SVM classification performance by re-sampling techniques. Lyhyaoui et al. (1999) conducted and improved the clustering for various samples, by selecting two samples with the closest distance from each cluster to establish the classifier.

This paper proposes a dual-process Sample Selection Support Vector Machine (SS-SVM) method for selecting informative genes form microarray dataset. We demonstrate the impact of SS-SVM on the feature selection method RFE-ISODATA, based on five cancer microarray datasets. According to the proportion of 3:1:1, the original dataset is randomly divided into the training set, validation set and independent test set. RFE-ISODATA is conducted on the samples selected by SS-SVM and all samples from the training set respectively. Experimental results show that SS-SVM method can effectively improve the classification and clustering capabilities of the informative genes selected by RFE-ISODATA.

The rest of this paper is organized as follows. Subsequently, the feature selection method of RFE-ISODATA is described, after which the dual-process sample selection method (SS-SVM) is proposed. This is followed by a presentation of the results of the feature selection experiments based on SS-SVM and an evaluation of the performance of the feature selection methods. Finally, this paper is concluded.

FEATURE SELECTION BASED ON FUZZY INTERACTIVE SELF-ORGANIZING DATA ALGORITHM

Fuzzy ISODATA is a kind of clustering algorithm with simple structure and high running speed (Bezdek, 1976; Marcelloni, 2003). The sample $X_i = \{x_{i1}, \ldots, x_{ij}, \ldots, x_{im}\}$ in training set belongs to **s** clusters and the k^{th} cluster center is represented as $V_k = \{v_{k1}, \ldots, v_{kj}, \ldots, v_{km}\}$. Membership u_{ki} of sample X_i which belongs to the k^{th} cluster is defined as (Bezdek, 1981; Marcelloni, 2003):

$$u_{ki} = \frac{1}{\sum_{t=1}^{s} \left(\frac{\|X_i - V_k\|}{\|X_i - V_t\|} \right)^{\frac{1}{r-1}}} \tag{1}$$

Membership u_{ik} regards the distance between the sample X_i and the cluster center V_k as the important indicator. Membership implies the relationship between features of sample and class of sample, so the features determine the membership value of sample to a certain class. The sensitivity formula of the j^{th} feature of samples to the membership (Liu et al., 2012) is defined as

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

$S(k,j)$ represents the sensitivity of the j^{th} feature to membership and reflects the contribution of the j^{th} feature to the k^{th} cluster. $S(j)$ can be regarded as the importance index of the feature in fuzzy ISODATA Clustering.

RFE-ISODATA method selects features based on the sensitivity index in the process of recursive feature elimination. In the process of feature sensitivity analysis, the “cluster” formed by the fuzzy ISODATA based on sample similarity reveals the underlying structure of the data. The discriminant function established by the features with high sensitivity has the high recognition ability.

As we know, the longer distance between sample and the center of other categories and the shorter distance between sample and the center of its own class will make the higher membership value of the samples. If the samples on the border of different classes are removed, spacing between different types of samples can be increased relatively and the remaining samples would have high membership value to their own classes in new round of fuzzy ISODATA clustering and the features with high sensitivity would carry more class information.

DUAL-PROCESS SAMPLE SELECTION BASED ON SVM (SS-SVM)

The support vector machine algorithm is a kind of machine learning algorithm developed by Vapnik based on the structural risk minimization principle and the statistical learning theory, which can obtain good generalization ability in the case of limited samples (Vapnic, 1998).

Let $X_i \in R^m$ be a sample of the training set X and $y_i \in \{+1, -1\}$ be a class label of X_i, that is, each sample X_i corresponds to a class indicator y_i. Linear discriminant function is given by $g(x) = \omega \cdot x + b$ and the hyperplane $g(x) = \omega \cdot x + b = 0$ can classify the training samples. Thus the margin, which is defined as the distance between the pair of parallel hyperplanes described by $\omega \cdot x + b = \pm 1$, is determined by ω which characterizes the direction of the hyperplane. To search for the maximum possible margin, quadratic programming problem is defined as below:

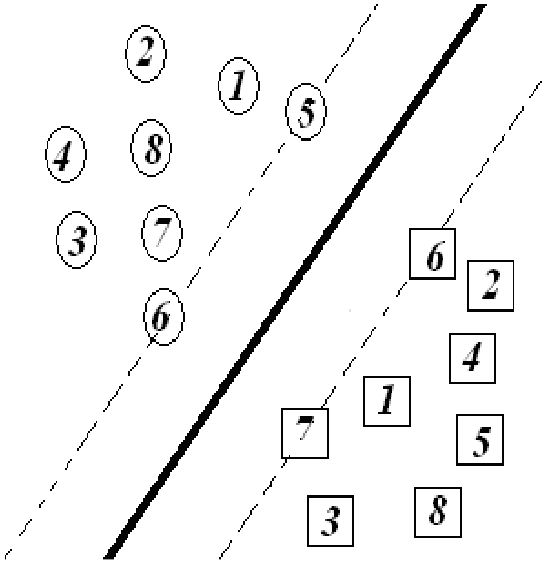

Figure 1. Distribution of two types of samples.

Minimize: $\Phi(\omega) = \frac{1}{2}\|\omega\|^2$

Subject to: $y_i(\omega^T \cdot X_i + b) \geq 1, i = 1,2,..., n$ (3)

Based on the method of Lagrange multipliers, the optimal SVM classification discriminant function is obtained as follows:

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{sv} \lambda_i^* y_i (X_i \cdot x) + b^*\right\} \quad (4)$$

The sample X_i with nonzero Lagrange multiplier λ_i^*, which lies on both hyperplanes, is called support vector sample (SV) (Vapnic, 1996). sv is the number of SVs. The SVs on the hyperplanes are only a small part in the training set, and are used as samples to define classification border.

As shown in Figure 1, 16 samples $(X_i, y_i), i = 1,...,16, X_i \in R^2$ belong to the positive and negative categories, respectively marked by the circular and square icons. The thick line in the middle refers to the optimal classification line and the dotted lines on both sides are 'support lines' of positive class and negative class. Six samples of the eight positive ones are at the upper left corner, the other two samples on the support line are the positive SVs. Six samples of the eight negative ones are at the lower right corner, the other two samples on the negative support line are the negative SVs.

Figure 1 indicates that if the SVs are removed, the spacing between the heterogeneous samples can be increased and cohesion of within-class samples can be enhanced relatively. Due to the fact that SVs only account for a small part of the training set, the removal of SVs will not affect the original information structure of dataset. In other words, if the samples excluding SVs are selected, not only the original class information of dataset can be retained, but also the spacing between different classes can be expanded, coupled with the shortening of inner-class distance. The dispersion between classes and compactness within-class are just the key factors to determine the class separability of feature in feature selection (Theodoridis and Koutroumbas, 1999). In view of this, this paper proposes a dual-process sample selection method (SS-SVM), which trains SVM on training set and selects samples other than SVs, for improving the performance of feature selection method.

EXPERIMENTAL RESULTS

The SS-SVM method is applied to feature selection method RFE-ISODATA and the impact of SS-SVM on RFE-ISODATA is studied via experiments. The flowchart of feature selection based on SS-SVM is shown in Figure 2. The SS-SVM method is added to recursive feature selection process: all samples in training set are used to train SVM, the samples selected by SS-SVM are used for feature selection. Based on these samples, features are sorted by RFE-ISODATA to generate candidate feature subsets.

To evaluate the class information of the candidate feature subsets reliably, classification and clustering tests are respectively done on the candidate feature subsets. The SVM and K nearest neighbor (KNN) classifiers trained by training set are used to identify the type of samples of the validation set for investigating the classification capability of the candidate feature subsets. Meanwhile, hierarchical clustering experiments are carried out on the validation set to check the clustering performance of the candidate feature subsets. The AUC(Area Under the receiver operating characteristic Curve) (Li et al., 2012; Provost and Fawcett, 1997) value and the correct rate of classification and clustering experiments on the validation set are used to build the objective function $Object(F)$ of the candidate feature subset F in the RFE process:

$$Object(F) = (AUC_{SVM}(F) + AUC_{KNN}(F) + AUC_{Cluster}(F) + right_{SVM}(F) + right_{KNN}(F) + right_{Cluster}(F))/6 \quad (5)$$

where AUC represents the AUC value of classification or clustering test, and $right$ stands for the right rate of classification or clustering test. The candidate feature subset with the highest value of the objective function is the optimal feature subset with the strongest classification and clustering performance in the validation test.

The classification and clustering performance of the optimal feature subset is further verified on the independent test set. The higher the AUC value and the right rate are, the stronger the classification and clustering ability of the optimal feature subset will be. SVs in SS-SVM algorithm will change along with the different features during the process of RFE. Despite of the shrinking range of feature selection, the scope of the sample selection remains all the samples of the training

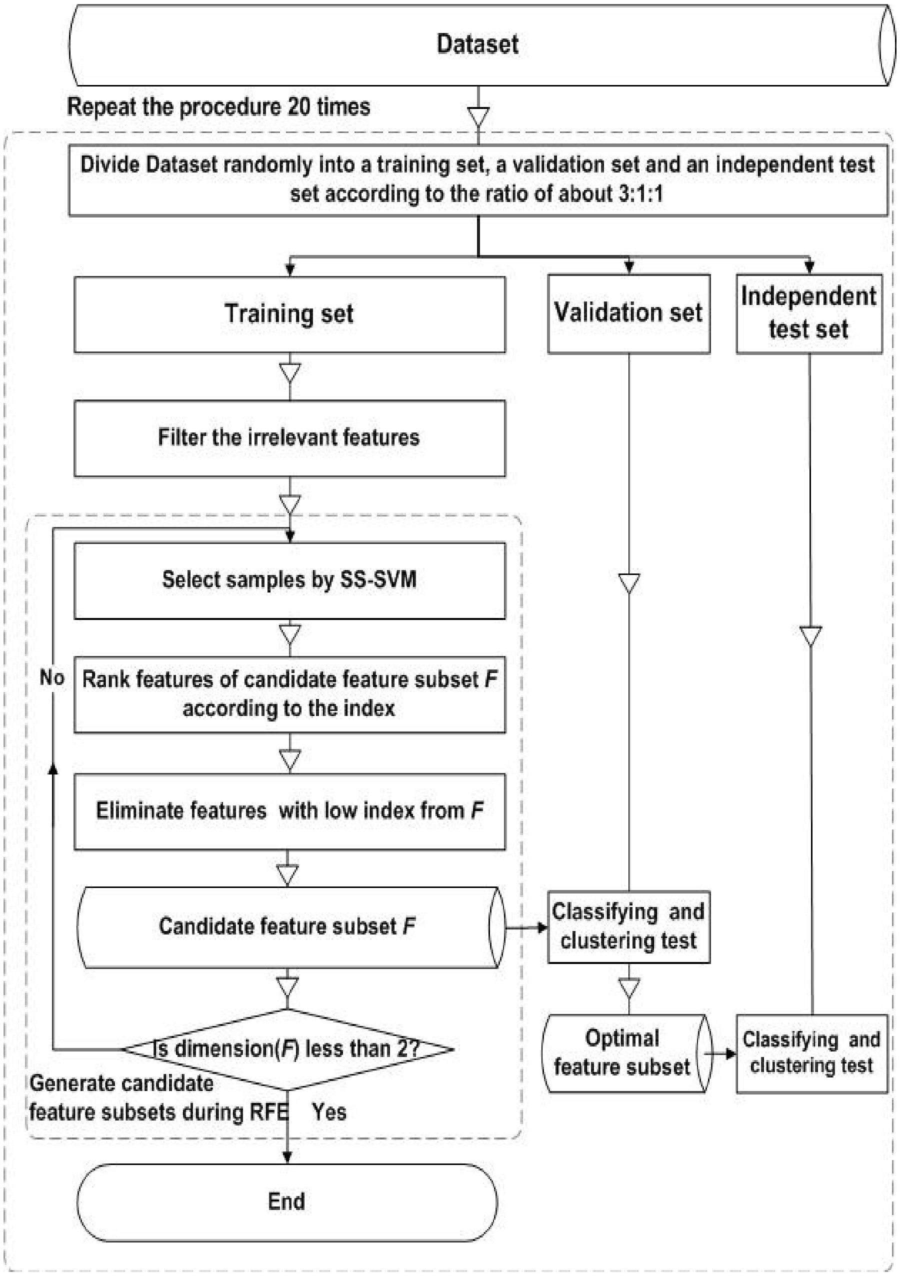
Dataset
Repeat the procedure 20 times
Divide Dataset randomly into a training set, a validation set and an independent test set according to the ratio of about 3:1:1
Training set
Validation set
Independent test set
Filter the irrelevant features
Select samples by SS-SVM
No
Rank features of candidate feature subset F according to the index
Eliminate features with low index from F
Candidate feature subset F
Classifying and clustering test
Is dimension(F) less than 2?
Generate candidate feature subsets during RFE
Yes
Optimal feature subset
Classifying and clustering test
End

Figure 2. The illustration of the process of feature selection based on SS-SVM.

set. Fixed scope of sample selection ensures the stability of the classification information of the selected features.

Microarray datasets

To verify objectively the effect of SS-SVM method on RFE-ISODATA, feature selection experiments are carried out on the 5 gene expression datasets (Alon et al., 1999; Goloub et al., 1999; Shipp et al., 2002; Singh et al., 2002; Zhan et al., 2002). As shown in Table 1, the number of genes in the datasets is greater than the number of samples, and different types of samples are unevenly distributed. In Table 1, the first column lists names of the datasets, the second column indicates the initial number of genes, and the third columns lists the number of samples of two different types. The samples of datasets are randomly assigned into the training set, validation set and independent test set by the proportion of 3:1:1 in the experiments.

On one hand, informative gene selection experiments based on RFE-ISODATA are carried out on the samples

Table 1. The datasets used for feature selection experiments.

Dataset	Features	Samples(+/-)	Training set/Validation set /Independent test set	Search scope of features
Colon	2000	62(40/22)	38/12/12	500
Acute Leukemia	7129	72(47/25)	44/14/14	100
Multiple myeloma	7129	105(74/31)	63/21/21	100
DLBCL	7129	77(58/19)	47/15/15	1000
Prostate	12600	102(52/50)	62/20/20	1000

selected by SS-SVM from the training set (RFE-ISODATA without SVs); on the other hand, informative gene selection experiments based on RFE-ISODATA are carried out on all samples of the same training set (RFE-ISODATA). To prevent the impact of uneven distribution of the samples on the feature selection, datasets are divided into three parts twenty times randomly. Consequently, both of feature selection methods are carried out on the three parts of the dataset each time. The statistical results of the feature selection experiments are used to evaluate the effect of SS-SVM on RFE-ISODATA.

Experimental configuration

The experiments are conducted with MATLAB on a PC with 2.2 GHz Intel Core 2 CPU and 2.0 GB.

The number of features in the datasets is greater than that of samples, with only part of features related to sample class. As a result, the irrelevant genes should be filtered before feature selection in order to reduce the search scope and the complexity of calculation. Bhattacharyya distance (Duda et al., 2001; Theodoridis and Koutroumba, 1999) between the heterogeneous samples in the training set is considered as criteria to filter the irrelevant genes. The number of filtered genes is determined through the filtering and classification tests on the five datasets. Table 1 lists the search scope for the next feature selection process.

For fuzzy ISODATA algorithm we set r = 2, s = 2, and ε = 0.0001. We set the kernel function of SVM as linear function and set 5 neighbors for the KNN algorithm. For the hierarchical clustering algorithm, we set the Euclidean distance as distance between pair-wised samples and construct the hierarchical cluster tree based on average distance.

Validation tests

Figure 3 shows the performance of classification and clustering in terms of objective function of 2 feature selection methods (RFE-ISODATA without SVs and RFE-ISODATA). The x-axis presents the number of the genes of the candidate feature subsets and the y-axis indicates the average value of the objective function in the validation tests during the 20 rounds of experiments.

The curve of objective function reflects the classification and clustering capabilities of candidate feature subsets in validation tests. The curves of “RFE-ISODATA without SVs” are higher than that of RFE-ISODATA, which indicates that the classification and clustering performance of the candidate feature subsets selected by the former is superior to the latter.

The curves of objective function indicate that SS-SVM method can improve the proportion of class information of the candidate feature subsets and is conductive for RFE-ISODATA.

Independent tests

The candidate feature subset with the highest objective function value is selected as the optimal feature subset. Genes in the optimal feature subset are considered as informative genes for cancer classification and diagnosis. To illustrate the effect of SS-SVM on feature selection methods, RFE-ISODATA is compared with “RFE-ISODATA without SVs” in terms of AUC value and right rate of classification and clustering in the independent tests.

Figure 4 illustrates the performance of the optimal feature subsets selected by the 2 feature selection methods from the 5 cancer microarray datasets respectively. The results indicate the mean and standard deviation of AUC value and right rate in the independent tests during the 20 rounds of feature selection experiments.

From Figure 4, we find the AUC value of classification and clustering tests of the optimal feature subsets selected by “RFE-ISODATA without SVs” is higher than that selected by RFE-ISODATA. Right rate of classification and clustering tests of the optimal feature subsets selected by “RFE-ISODATA without SVs” is higher than that selected by RFE-ISODATA, except the SVM classification performance on the Prostate dataset. Comparison of the independent test result in Table 2 indicates that the classification and clustering performance of the informative genes selected by “RFE-

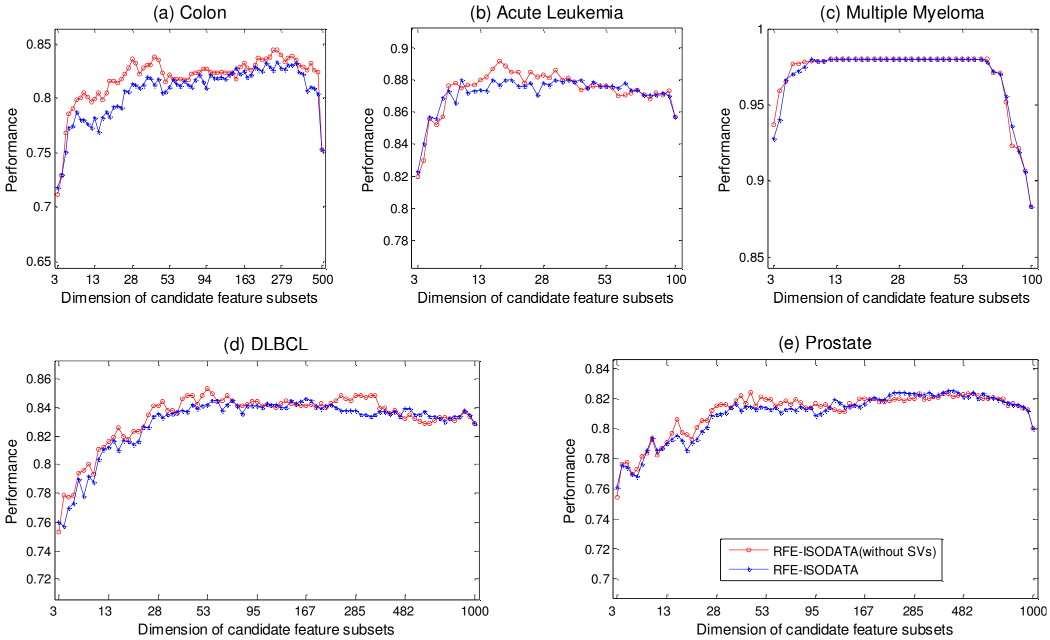

Figure 3. Objective function curve of 2 feature selection methods on 5 microarray datasets.

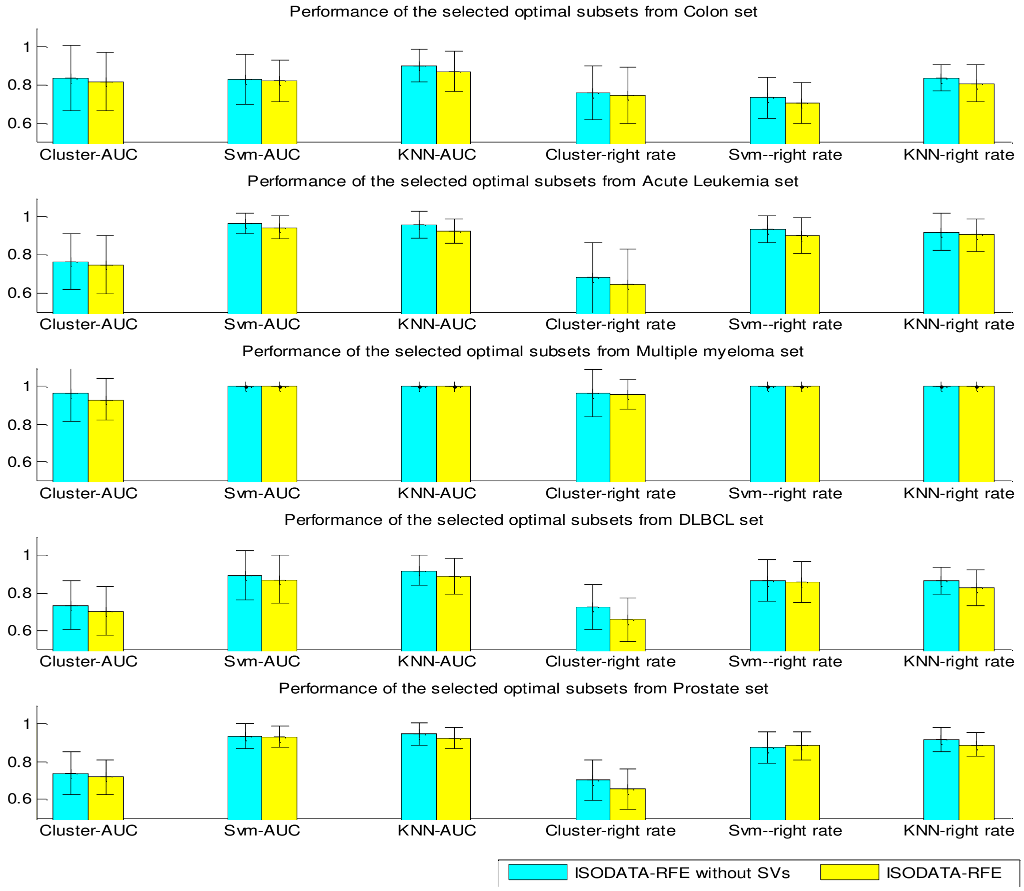

Figure 4. Performance of the optimal feature subsets selected by the 2 feature selection methods in the independent tests.

Table 2. Performance of the selected optimal feature subsets in the independent tests.

Microarray dataset	Feature selection method	Dimension	AUC			Right rate		
			SVM	KNN	Hierarchical clustering	SVM	KNN	Hierarchical clustering
Colon	RFE-ISODATA without SVs	18.35	0.839±0.133	0.903±0.088	0.837±0.172	0.733±0.085	0.838±0.069	0.758±0.145
	RFE-ISODATA	18.35	0.821±0.109	0.871±0.106	0.816±0.155	0.704±0.106	0.808±0.098	0.746±0.149
Acute Leukemia	RFE-ISODATA without SVs	8.95	0.966±0.052	0.962±0.070	0.762±0.148	0.946±0.061	0.932±0.047	0.725±0.160
	RFE-ISODATA	9.00	0.943±0.062	0.941±0.067	0.726±0.156	0.914±0.082	0.921±0.065	0.711±0.157
Multiple Myeloma	RFE-ISODATA without SVs	4.75	1.000±0.000	1.000±0.000	0.967±0.151	1.000±0.000	1.000±0.000	0.967±0.078
	RFE-ISODATA	4.00	1.000±0.000	1.000±0.000	0.932±0.111	1.000±0.000	1.000±0.000	0.960±0.053
DLBCL	RFE-ISODATA without SVs	20	0.895±0.131	0.903±0.080	0.733±0.130	0.863±0.111	0.863±0.073	0.723±0.119
	RFE-ISODATA	19.55	0.875±0.128	0.888±0.094	0.703±0.128	0.860±0.110	0.823±0.098	0.657±0.119
Prostate	RFE-ISODATA without SVs	21.8	0.936±0.064	0.949±0.061	0.738±0.112	0.873±0.082	0.918±0.067	0.703±0.107
	RFE-ISODATA	17.05	0.932±0.058	0.926±0.056	0.717±0.093	0.883±0.076	0.890±0.062	0.653±0.110

ISODATA without SVs" is superior to that selected by RFE-ISODATA. It shows that the class information carried by the informative genes selected by "RFE-ISODATA without SVs" is much more than that selected by RFE-ISODATA. This indicates that SS-SVM method can effectively improve the performance of RFE-ISODATA on the 5 microarray datasets.

Integrating the results of 20 rounds of feature selection experiments by the 2 feature selection methods on the 5 cancer microarray datasets, it verifies that the removal of SVs can expand the distance between heterogeneous classes, shrink the sample dispersion within-cluster in fuzzy ISODATA clustering, increase the sensitivity of feature to sample membership and enhance the classification performance of the selected informative genes.

Results of the experiments show SS-SVM can improve the recognition ability and clustering performance of the selected informative genes by RFE-ISODATA. It means the application of SS-SVM on feature selection is useful to cancer diagnosis and findings of subtype of cancer.

Conclusions

This paper proposes a new method of dual-process sample selection based on SVM (SS-SVM) and studies the impact of SS-SVM on feature selection method RFE-ISODATA. In this paper, we show that SS-SVM is able to relatively contract the dispersion within class and extend the distance between the classes by removing SVs, and thus improve the clustering quality of fuzzy ISODATA and the feature selection performance of RFE-ISODATA.

Informative gene selection experiments based on 5 microarray datasets show that SS-SVM method can effectively improve the performance of RFE-ISODATA algorithms. SS-SVM combined with RFE-ISODATA achieved high clustering performance on the independent test sets, implying this combined method has the potential to identify cancer subtypes. Thus, SS-SVM could have its application potential in cancer diagnosis and drug response.

As the 5 cancer microarray datasets contain only dozens of samples, we will take further feature selection tests based on SS-SVM on the datasets with more samples.

This paper develops the sample selection method (SS-SVM) which helps feature selection method to identify the key genes with abundant class information from the gene expression datasets. Experimental results prove that SS- SVM plays an important role in improving the performance of feature selection method based on clustering model. In the future, we will study the impact of SS-SVM on the feature selection methods based on the classification model.

ACKNOWLEDGEMENTS

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

Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y (2010). Robust biomaker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26:392-298.

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

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS USA 96:6745-6750.

Bezdek JC (1976). Physical interpretation of fuzzy ISODATA. EEE SMC, SMC-6 pp. 387-390.

Bezdek JC (1981).Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Duda RO, Hart PE, Stork DG (2001). Pattern Classification. 2nd ed, John Wiley & Sons.

Goloub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, BloomÞeld CD, Lander ES (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Sciences 286:531-537.

Guyon I, Elisseeff A (2003). An Introduction to Variable and Feature Selection, J. Mach. Learn. Res. 3:1157-1182.

Guyon I, Weston J, Barnhill S, Vapnik V (2000). Gene selection for cancer classification using support vector machines. Mach. Learn. 46(13):389-242.

Jensen R, Shen Q (2009). Feature Sejhiojlection for Aiding Glass Forensic evidence analysis, Intel. Data Anal. 13(5):703-723.

Lan L, Vucetic S (2011). Improving accuracy of microarray classification by a simple multi-task feature selection filter. Int. J. Data Min. Bioinforma. 5(2):189-208.

Li Y-X, Ji S, Kumar S, Ye J, Zhou, Z-H (2012). Drosophila gene expression pattern annotation through multi-instance multi-label learning. ACM/IEEE Trans. Comput. Biol. Bioinforma. 9(1):98-112.

Liu QJ, Zhao ZM, Li Y-X, Li YY (2012). Feature Selection Based on Sensitivity Analysis of Fuzzy ISODATA. Neurocomputing 85:29-37.

Lyhyaoui A, Martínez M, Mora I, Vázquez M, SanchoJ, R Figueiras-Vidal A (1999). Sample selection via clustering to construct support vector-like classifiers. IEEE Trans. Neural Netw. 10(6):1474-1481.

Marcelloni F (2003). Feature selection based on a modified fuzzy C-means algorithm with supervision. Inform. Sci. 151:201-226.

Provost F, Fawcett T (1997). Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: Proc. Third Int. Conf. on Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA, pp. 43-48.

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar R CT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister A, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 8(1):68-74.

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1(2):203-209.

Tang YC, Zhang YQ, Clhawla NV, Krasser S (2009). SVMs modeling for highly imbalanced classification, IEEE Trans. Syst., Man, Cybern. - Part B Cybern. 39:281-288.

Tang YC, Zhang YQ, Huang Z, Hu XH, Zhao YC (2008). Recursive fuzzy granulation for gene subsets extraction cancer classification, IEEE Trans. Inform. Technol. Biomed. 12(6):723-730.

Theodoridis S, Koutroumbas K (1999). Pattern Recognition, Academic Press, New York.

Uncua Ö, Türkşenb IB (2007). A novel feature selection approach: Combining feature wrappers and filters. Inf. Sci. 177(2):449-466.

Vapnic VN (1998). Statistical Learning Theory. John Wiley and Sons, New York.

Vapnic VN (1996). The Nature of Statistical Learning Theory. Springer, New York.

Zhan F, Hardin J, Kordsmeier B, Bumm K, Zheng MZ, Tian E, Sanderson R, Yang Y, Wilson C, Zangari M, Anaissie E, Morris C, Muwalla F, Rhee FV, Fassas A, Crowley J, Tricot G, Barlogie B, John Shaughnessy JJ (2002). Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. Blood 99:1745-1757.