BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 11(1)

A novel feature extraction approach for microarray data based on multi-algorithm fusion

Zhu Jiang[1,2]* & Rong Xu[1]

[1]State Key Laboratory of Astronautic Dynamics, Xi'an Satellite Control Center, Xi'an, China; [2]Key Laboratory of Fluid and Power Machinery, Ministry of Education, Xihua University, Chengdu, China; Zhu Jiang - Email: hill5525@163.com; *Corresponding author

Received December 09, 2014; Revised December 31, 2014; Accepted January 08, 2015; Published January 30, 2015

Abstract:
Feature extraction is one of the most important and effective method to reduce dimension in data mining, with emerging of high dimensional data such as microarray gene expression data. Feature extraction for gene selection, mainly serves two purposes. One is to identify certain disease-related genes. The other is to find a compact set of discriminative genes to build a pattern classifier with reduced complexity and improved generalization capabilities. Depending on the purpose of gene selection, two types of feature extraction algorithms including ranking-based feature extraction and set-based feature extraction are employed in microarray gene expression data analysis. In ranking-based feature extraction, features are evaluated on an individual basis, without considering inter-relationship between features in general, while set-based feature extraction evaluates features based on their role in a feature set by taking into account dependency between features. Just as learning methods, feature extraction has a problem in its generalization ability, which is robustness. However, the issue of robustness is often overlooked in feature extraction. In order to improve the accuracy and robustness of feature extraction for microarray data, a novel approach based on multi-algorithm fusion is proposed. By fusing different types of feature extraction algorithms to select the feature from the samples set, the proposed approach is able to improve feature extraction performance. The new approach is tested against gene expression dataset including *Colon cancer data*, *CNS data*, *DLBCL data*, and *Leukemia data*. The testing results show that the performance of this algorithm is better than existing solutions.

Keywords: feature extraction; robustness; microarray data; multi-algorithm fusion

Background:

Feature extraction is one of the key issues in the field of data mining. Researchers have realized that in order to use data mining tools effectively, data preprocessing is essential to successful data mining. Feature extraction is one of the important and frequently used techniques in data preprocessing. It not only can eliminate information redundant, improve the classification efficiency and accelerate the computational speed, but also can reduce the complexity of the classifier and the error rate of classification **[1]**. Feature extraction algorithm can be classified into three fundamentally approaches: wrapper, filter and embedded. Wrapper model evaluates the subset of selected features by using criteria based on the results of learning algorithms, while filter methods depends on intrinsic characteristics of the data to select feature subsets without involving and mining methods **[2]**. These methods are often limited in use because they require a long computational time. To take advantage of these two algorithms, the embedded method is proposed **[3]**.

For all feature extraction algorithms, many real applications, such as computer vision, microarray technology and visual recognition, involve microarray data. Robustness of feature extraction for these data also gets its attentions in recent years. Gulgezen *et al.* **[4]** studied the stability and classification accuracy of Minimum Redundancy Maximum Relevance-based feature extraction. An entropy-based measure for stability assessment was developed by Krizek *et al* **[5]**. However, research shows that the extraction results of most feature extractions are very sensitive to the changes of training sets. That is to say, these algorithms have poor robustness **[6]**. This problem is particularly obvious for microarray data set. Even if the training data set is slightly changed, the obtained optimum feature subset will have large difference and the performance of the classification model will change greatly. Therefore, to improve the credibility of classification performance, we need to choose the feature extraction algorithm with a high robustness.

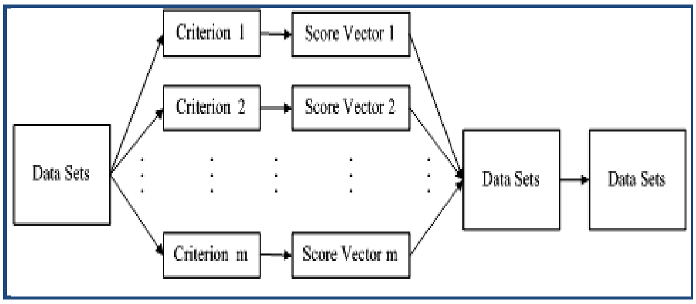

Figure 1: Score-based multi-algorithm fusion.

In previous studies, it is well known that combining or integrating multiple classifiers, especially uncorrelated weak ones, could greatly improve the classification performance **[7],** but studies for the fusion of various feature extraction algorithms are a few. Marina Skurichina believes that there may be useful information in the unselected features after the feature extraction. The omission of these features may lead to the poor performance of feature extraction and pattern recognition. So it is suggested to use the method of fusion to utilize the useful information in the neglected features **[8].** As we all know that feature subsets produced by different feature extraction algorithm may show complementary, and fusion of multiple algorithms utilizes the search abilities of each algorithm to get closer to a global optimal solution. Thus a fusion of the feature subsets may produce a better representation in feature space.

A lot of feature extraction algorithms have been proposed in the literature. But not all feature extraction algorithms can be fused. If two extraction algorithms are similar, the fusion of them can not improve the stability of the extraction algorithm greatly **[9].** Therefore, the diversity must be considered while choosing the feature extraction algorithms. Different types of feature extraction can complement each other and avoid overlapping. Obviously, it is not necessary to fuse all feature extraction algorithms, which is also impossible. To simplify the calculation process and reduce the amount of computation while maintaining the diversity of extraction algorithms, in this study, Fisher Ratio, Absolute Weight of SVM (AW-SVM) and Polynomial Support Vector Machine (PSVM) were used to fuse in this paper. Fisher Ratio **[9]** is one of the basic methods in filter mode of feature extraction. It is a univariate filter method evaluating each feature individually. Its estimation standards are directly obtained from the data set. It has the characteristics of small calculation cost, high efficiency, etc.. AW-SVM **[10]** is an embedded method that ranks features based on their corresponding coefficients in the SVM classifier. PSVM is a wrapper method based on statistical learning theory **[11].** It has powerful fault tolerance and generalization abilities. Studies have shown that the generalization ability of PSVM will not reduce with the increasing of the order. It overcomes the problems of over learning, lack of learning, local minimum value and dimension disaster of traditional machine learning.

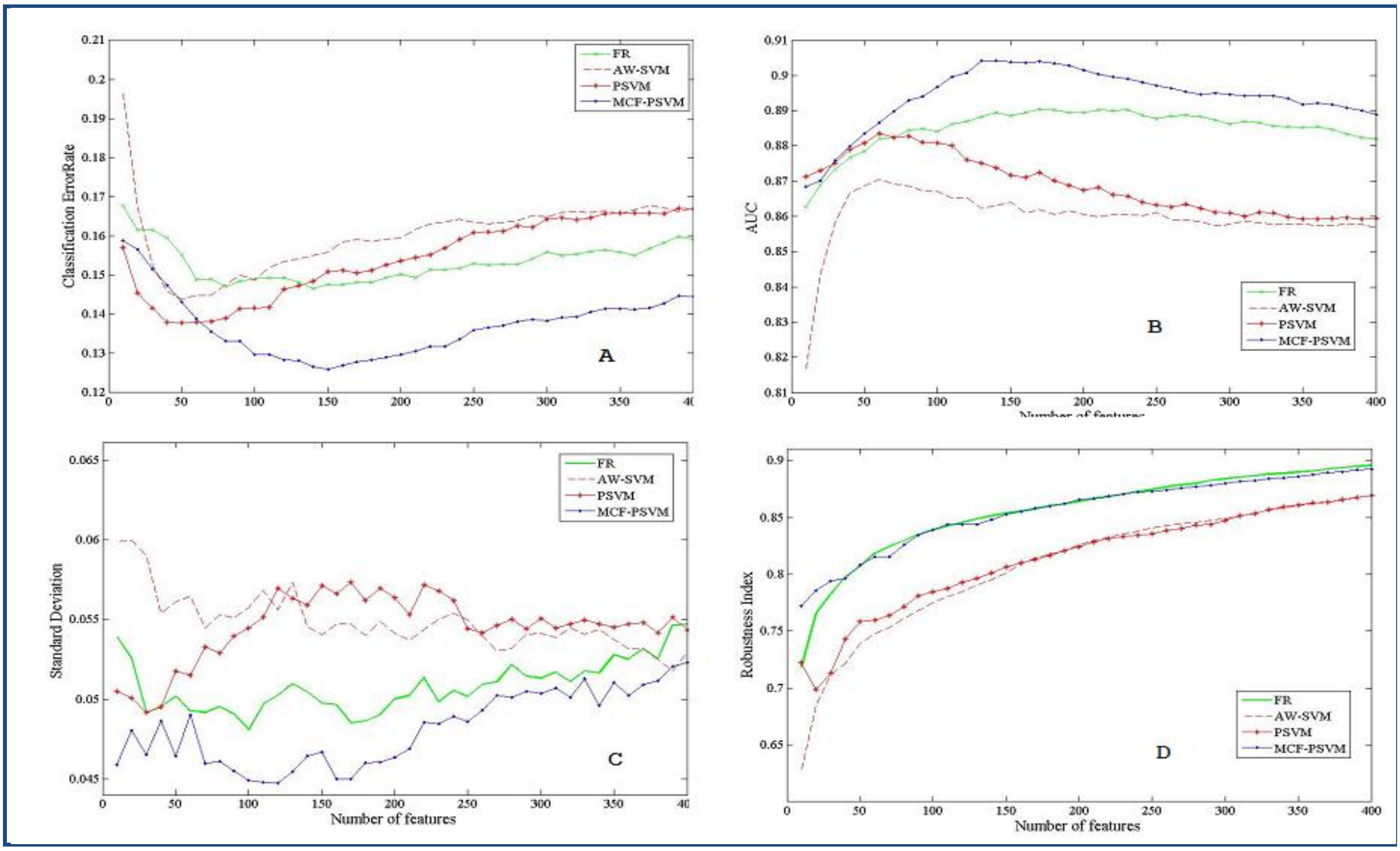

Figure 2: Performance comparisons on colon data: **A)** Classification Error; **B**) AUC; **C)** Standard Deviation of Error Estimation; **D)** Feature Robustness.

By considering all these factors and based on the idea of fusion, a novel feature extraction method, polynomial support vector machine based on multi-algorithm fusion (MAF-PSVM), was put forward in this paper.

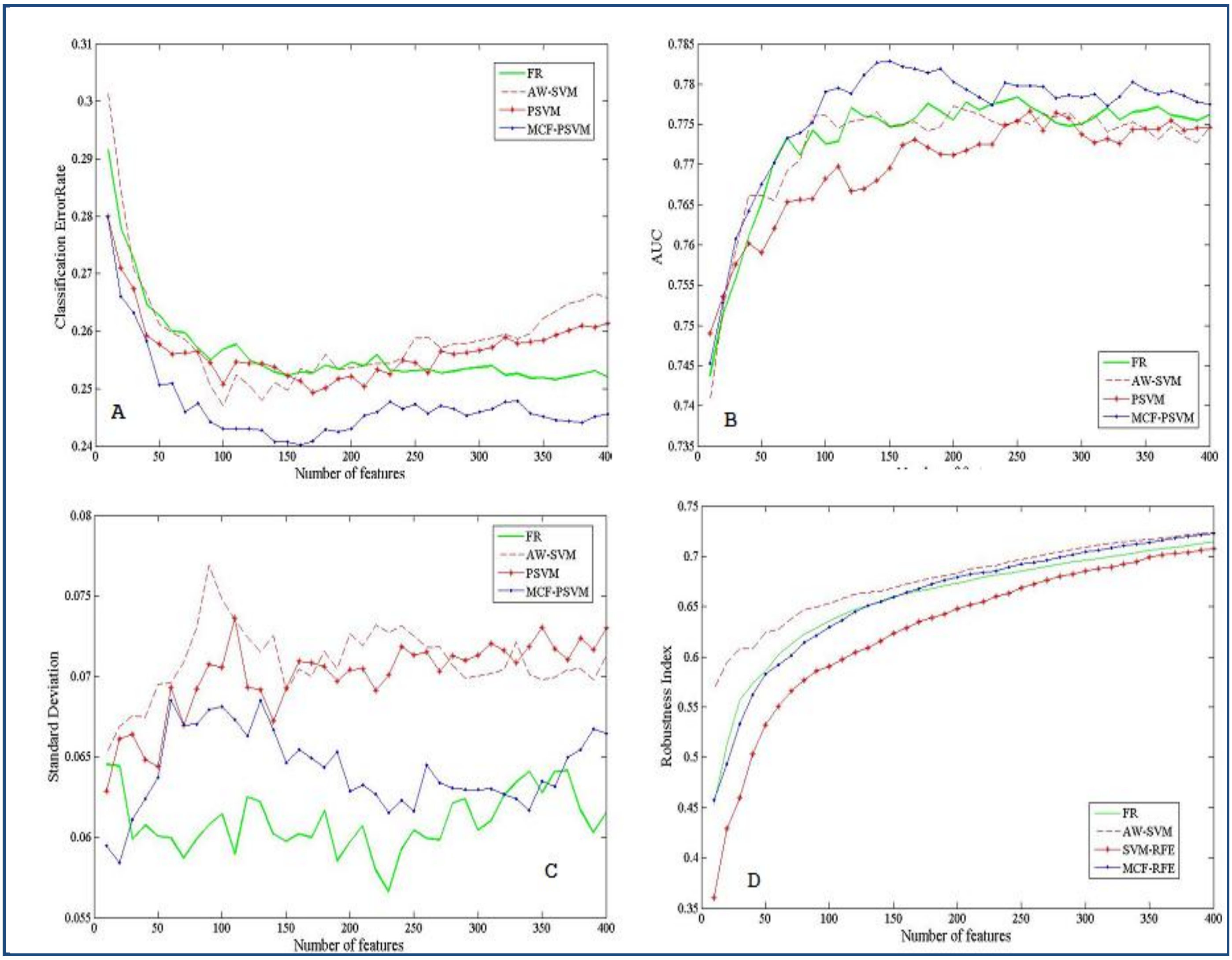

Figure 3: Performance comparisons on CNS data: **A)** Classification Error; **B)** AUC; **C)** Standard Deviation of Error Estimation; **D)** Feature Robustness.

Methodology:

Polynomial Support Vector Machine based on Multi-Algorithm Fusion

The specific implementation steps of MAF-PSVM are introduced in detail.

Step one, initialize the sample data set. Training samples are clustered into K classes by k-means **[12]** (The value of k in this paper was 8 by training.), different penalty factors are used for each class.

Step two, fuse feature extraction algorithms (in this paper, the algorithms were Fisher Ratio, AW-SVM and PSVM (How to select appropriate kernel functions for different applications has been a difficult problem. Studies have shown that linear kernel function is suitable for the case of linearly separable data, so polynomial kernel was selected. The other kernel functions can be further studied in the future for us. The generalization ability of polynomial kernel functions is different. In most cases, the performance of the classifier can reach to optimum while the parameter *d* is taken as 1**[13].** Therefore, the value of order *d* of the polynomial kernel was taken as 1)) and conduct feature extraction for clustered samples.

The following contents introduce the fusion method used by this paper:

In our study, we used score-based multi-algorithm fusion methods. Firstly, a score vector containing scores of all features was produced by each basis criterion. Secondly, a score combination algorithm was used to aggregate the multiple score vectors into one consensus score vector. Finally, a feature ranking procedure was employed to rank the features based on their consensus scores. The score-based multi-algorithm fusion procedure is illustrated in **Figure 1.**

In score aggregating, the scores produced by different basis criteria will be comparable. It is essential to ensure that score normalization should be done before score combination is performed. The scores are normalized to the range of **[0, 1]** in this study. Assume u*ᵢ* is the score vector produced by basis criterion *i*, the score normalization is performed as follows:

$$u_i^{'} = \frac{u_i - u_{i\,\min}}{u_{i\,\max} - u_{i\,\min}} \qquad (1)$$

where $u_{i\,\min}$ and $u_{i\,\max}$ are the minimum and maximum values in vector u_i.

For all the basis criteria, it is assumed that the larger the score, the better the feature. A score combination method will be used as following:

$$u = \frac{1}{m}\sum_{i=1}^{m} u_i^{'} \qquad (2)$$

where *m* is the number of basis criteria used in fusion.

Step three, feature extraction results of the step two will be used to train the PSVM classifier.

If the numbers of negative class points and positive class points have a large difference in the training data set, and if applying the same penalty parameter *C* to the set of positive class points and the set of negative class points, it means that the one with more class points will get more attention. However, we hope that penalties to the positive point and negative point are not the same. Accordingly, for properly selected parameter *C*, $C_+ = \frac{l_-}{l_+ + l_-}C$ and $C_- = \frac{l_+}{l_+ + l_-}C$. Where, l_+ and l_- are respectively the number of positive class training points and negative class training points. C_+ is the penalty parameter of positive class point. Comparatively, C_- is the penalty parameter of negative class point.

The PSVM classifier was built as following:

$$\begin{cases} \max W(\alpha) = \frac{1}{2}\sum_{i=1}^{l}\alpha_i - \sum_{i,j=1}^{l}\alpha_i\alpha_j(x_i \cdot x_j + 1)^d y_i y_j \\ \text{s.t.}\sum_{i=1}^{l}\alpha_i y_i = 0, \\ 0 \le \alpha_i \le C_{class1}, Class \quad Index = class1; \\ 0 \le \alpha_i \le C_{class2}, Class \quad Index = class2; \\ \cdots\cdots\cdots\cdots \\ 0 \le \alpha_i \le C_{classN}, Class \quad Index = classN; \\ C_i = \frac{l_1 +, \cdots, l_{i-1}, l_{i+1} \cdots, + l_n}{l_1 +, \cdots, + l_n} \end{cases}$$

(3)

Where, α represents the Lagrange multiplier. *class*1, …, *classN* represent the categories after clustering. *Class Index* represents the mark of class. l_1, …, l_n represent the number of sample points in each class, and C_i represents the penalty factor of each class.

Step four, regress the sample data set by trained classifier, and remove features with the minimal correlation. Sample set will be updated. In this study, in order to give a more general and precise measure of the similarity between two feature subsets, the following similarity index *JC* (∈ *[0,1]*) that takes in account the correlations between the different features of two feature subsets will be used.

$$JC_i(k) = \frac{|S_i \cap S_0| + SC_i}{k} \qquad (4)$$

Where S_i and S_0 are feature subsets selected using the *i*th batch of re-sampled data and the full data respectively. SC_i is the sum of absolute correlation values between the dissimilar features form S_i and S_0. It will be computed using the greedy search algorithm.

Step five, see if the coding is over, that is, original feature set *S*= [1, 2, …, *n*] is null. If so, end the iteration, otherwise repeat step two to four until achieving the feature extraction.

Measurement of feature extraction robustness

The definition of robustness of feature extraction is the sensitivity degree of the results of feature extraction algorithm to the changes of training set. According to this definition, the measurement of feature extraction robustness is measuring the similarity among the optimal feature subsets selected by the algorithm. The overall robustness of the algorithm can be calculated as **[14]:**

$$S_{tot} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S(s_i, s_j)}{k(k-1)} \qquad (5)$$

Where, s_i represents the results of feature extraction from training set No. $i(1 \le i \le k)$. $S(s_i, s_j)$ represents a similarity measure between two feature extraction results s_i and s_j. At present, there are many types of similarity measuring methods according to different representation ways of feature extraction results. The commonly used set method **[14]** was selected by this study. In this method, the robustness is measured by Tanimoto distance **[12].** Tanimoto distance is used to measure the coincidence degree of elements between two feature subsets:

$$S_s(s_i, s_j) = 1 - \frac{|s_i| + |s_j| - 2|s_i \cap s_j|}{|s_i| + |\sigma_j| - |s_i \cap s_j|}$$

$$= \frac{|s_i \cap s_j|}{|s_i \cup s_j|} = \frac{\sum_i \mathrm{I}(s_i^l = s_j^l = 1)}{\sum_j \mathrm{I}(s_i^l + s_j^l > 0)}$$

If the parameter is true, the function $\mathbf{I}(\cdot)$ returns as 1; otherwise it returns as 0. The value interval of S_s is [0, 1]. 0 means that the intersection of these two sets is an empty set, and all elements are different. 1 means that these two sets are exactly the same, and all elements are the same. The sizes of sets measured by Tanimoto distance can be the same or different.

Results:

Colon cancer data, *CNS data*, *DLBCL data*, and *Leukemia data* **[15]** were separately adopted for the simulation test. Aiming at these microarray data sets, performance evaluation of feature extraction was conducted for the method proposed in this paper, Fisher Ratio, AW-SVM and PSVM in four aspects of identification error, AUC values, standard deviation and robustness. The results were shown in **Figure 2 - Figure 5.**

It can be seen from the simulation results (in the **Figure 2a, Figure 3a, Figure 4a, & Figure 5a)** that the accuracy of feature identification for the method proposed in this paper is better than the other three methods. Taking **Figure 2a** as an example, MAF-PSVM realizes the minimum identification error by only extracting 150 features, and the identification error at this time is 12.96%. However, the identification errors of Fisher Ratio, AW-SVM and PSVM when extracting 150 features are 14.70%, 15.63% and 15.17%, separately.

The area (AUC) in the ROC curve is usually used to measure the classification performance. The larger the AUC values are, the better the classification performance will be. So AUC values are used to evaluate the classification performance of several feature extraction methods during the simulation test. The AUC values of these four methods are shown in **Figure 2b, Figure 3b, Figure 4b,** & **Figure 5b.** It can be seen from these results and the identification errors of all methods shown in **Figure 2a, Figure 3a, Figure 4a & Figure 5a** that when extracting 150 features, the AUC values of the method proposed in this paper are better than the other three methods when colon data and CNS data were used. When DLBCL data and Leuk data are used to test, the AUC values of MAF-PSVM are close to the result of FR. It also indicates that the classification performance of MAF-PSVM is better than the other three methods while achieving the most accurate extraction of features.

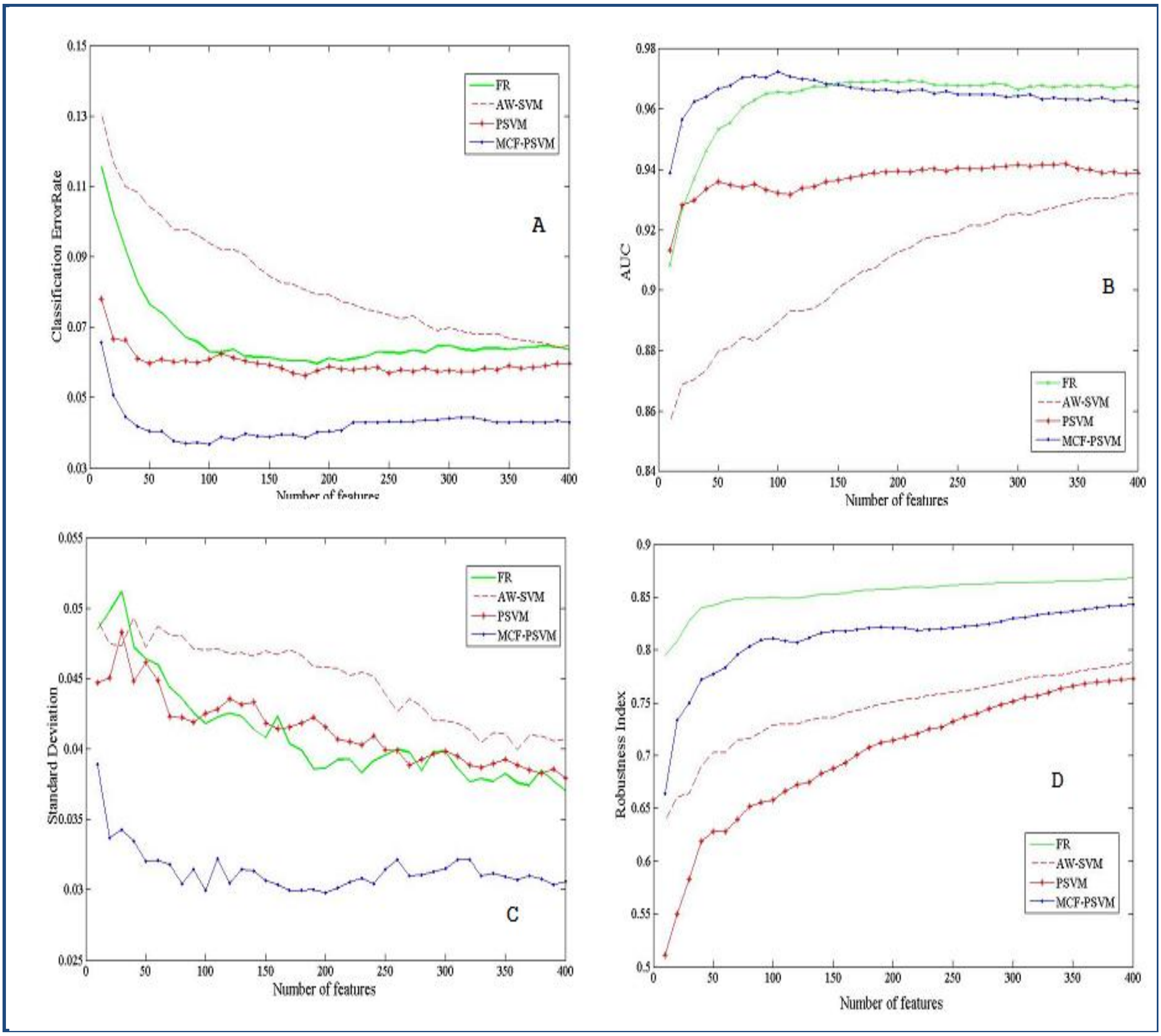

Figure 4: Performance comparisons on DLBCL data**: A)** Classification Error; **B)**AUC; **C)** Standard Deviation of Error Estimation**; D**) Feature Robustness.

The standard deviations of all methods are shown in **Figure 2c, Figure 3c, Figure 4c, and Figure 5c**. By analyzing the simulation results, it is known that the performance of the method proposed in this paper is better than the other three feature extraction methods when Colon and DLBCL data are used. For example, on colon data **(Figure 2c),** when extracting 150 features, the standard deviation of MAF-PSVM is only 0.0456. Its identification accuracy is second only to the Fisher Ratio, which standard deviation is 0.050 at this time; the standard deviation of AW-SVM is 0.547, and that of PSVM is the largest, which reaches to 0.0561.

The robustness results of all methods during the feature extraction process are shown in **Figure 2d, Figure 3d, Figure 4d, Figure 5d.** By analyzing the simulation results, it is known that the stability of the method proposed in this paper does not perform the best. This is because the method proposed in this paper is an embedded feature extraction method. Compared with other algorithms, it fully considers the dependence among features during the feature extraction process. The results of this treatment are that the feature can be extracted more accurately and the identification of patterns can be realized. The simulation results of the estimated classification error, AUC, and the standard deviation of error estimation can fully confirm this.

It is worth mentioning that while evaluating the performance of a feature extraction method, we need to comprehensively consider the accuracy, efficiency and stability of feature identification of the method. Classification performance should be the first consideration because a classification-ineffective extraction result does not make any sense. Based on this and the above simulation analysis results, we can conclude that the comprehensive performance of MAF-PSVM proposed in this paper is better than the other three methods during feature extraction of microarray data.

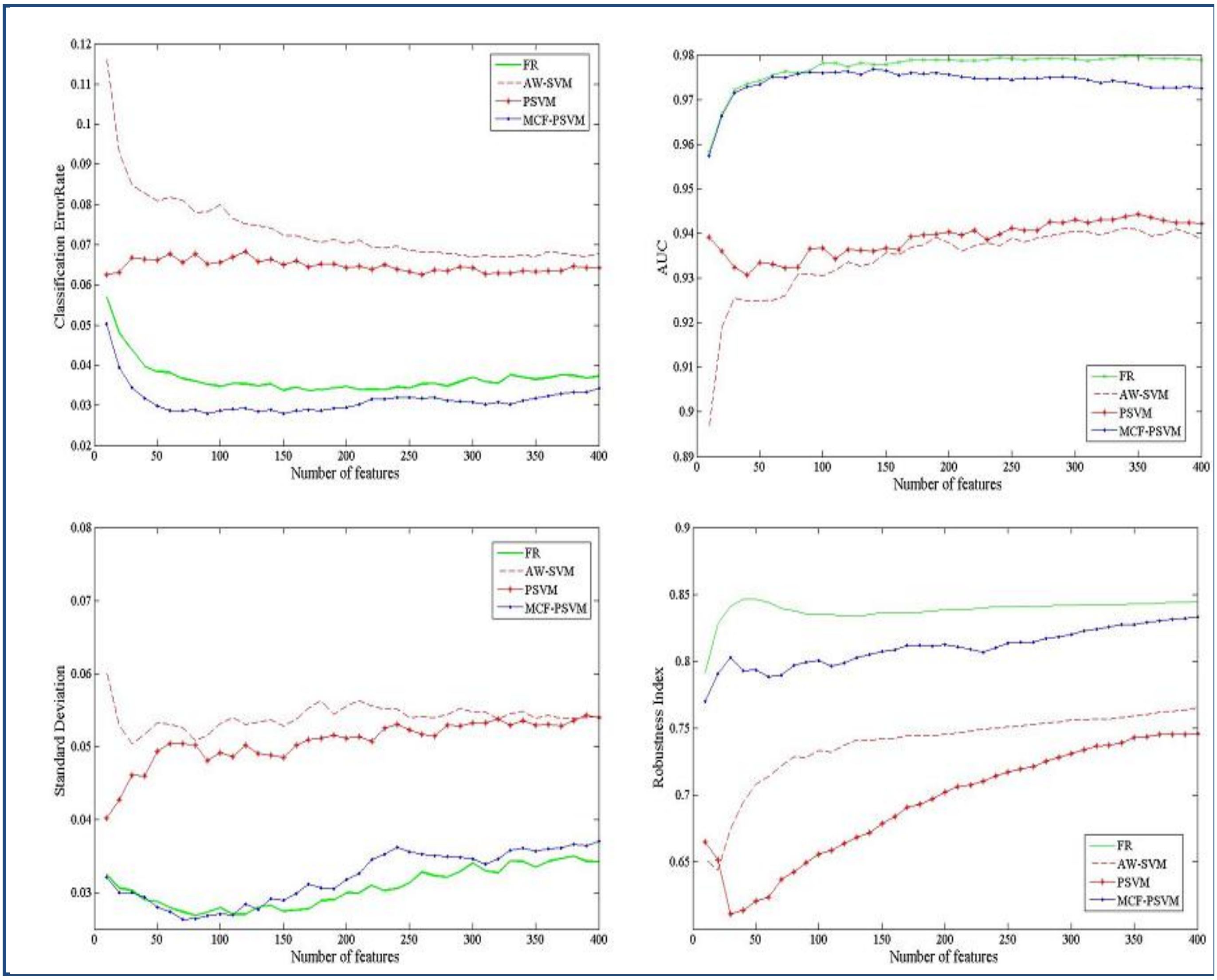

Figure 5: Performance comparisons on Leukemia data: **A)** Classification Error; **B)** AUC; **C)** Standard Deviation of Error Estimation; **D)** Feature Robustness.

Conclusion:

The feature extraction for microarray data was discussed and analyzed in this paper. According to the idea of clustering and information fusion, a novel feature extraction method, polynomial support vector machine based on multi-algorithm fusion (MAF-PSVM), was put forward. The simulation results of measured data show that the identification error, AUC values, standard deviation of error estimation of the method proposed in this paper are better than Fisher Ratio, AW-SVM and PSVM. It was found out while analyzing and comparing the robustness of feature extraction of all methods that the method proposed in this paper does not perform the best. This is because it fully considered the dependence among features. The new method builds a compromise between the accuracy of feature identification and the stability of feature extraction. For the performance evaluation of a kind of feature extraction method, both the stability of feature extraction and the identification performance (such as identification accuracy and efficiency) shall be considered.

Therefore, the comprehensive performance of the MAF-PSVM is better than other methods. This method is more suitable for the feature extraction of microarray data.

Acknowledgement:
Funding: the research is supported by key laboratory of fluid and power machinery, ministry of education.xihua university (szjj2014-038) and laboratory of pattern recognition and artificial intelligence.

References:

[1] Langley P, *Artificial Intelligence*. 1997 **97**: 245
[2] Guyon I *et al. Journal of Machine Learning Research*. 2003 **3:** 1157
[3] Maldonado SR *et al. Information Sciences*. 2011 **181**: 115
[4] Gulgezen GZ *et al. Book series in Machine Learning and Knowledge Discovery in Databases*. 2009 **781**: 455
[5] Krizek PJ *et al. Lecture Notes in Computer Science*. 2007 **4673**: 929
[6] Davis CA *et al. Bioinformatics* 2006 **22:** 2356 [PMID: 16882647]
[7] Yang F & Mao KZ, *IEEE/ACM Trans Comput Biol Bioinform.* 2011 **8**: 1080 [PMID: 21566255]
[8] Skurichina M *et al. Multiple Classifier Systems*. 2005 **3541**: 165
[9] Fung ES & *Nq MK, Bioinformation* 2007 **2**: 230 [PMID: 18305833]
[10] Ying-Xin L *et al. Chinese Journal of Computers*. 2006 **23**: 324
[11] Jiang H & Ching WK, *Bioinformation* 2011 **7:** 257 [PMID: 22125395]
[12] Ge Y & Sealfon SC, *Bioinformatics* 2012 **28**: 2052 [PMID: 22595209]
[13] Dwork C *et al. In World Wide Web*. 2001: 613-622
[14] Kalousis A *et al. Knowledge and Information System*. 2007 **12**: 95
[15] http://www-genome.wi.mit.edu

Edited by P Kangueane

Citation: **Jiang & Xu,** Bioinformation 11(1): 027-033 (2015)