African Journal of Biotechnology Vol. 10(74), pp. 16720-16731, 23 November, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.2273
ISSN 1684–5315 © 2011 Academic Journals

Full Length Research Paper

A novel stepwise support vector machine (SVM) method based on optimal feature combination for predicting miRNA precursors

Limei Wang[1#], Jin Li[1#], Rongsheng Zhu[1,2], Liangde Xu[1], Ying He[1], Ruijie Zhang[1]* and Shaoqi Rao[1]*

[1]Department of Bioinformatics and Computer, Harbin Medical University, Harbin 150081, China.
[2]College of Science, Northeast Agricultural University, Harbin 150030, China.

Accepted 24 October, 2011

MicroRNAs (miRNAs) are a class of non-coding RNAs that are produced from miRNA precursors (pre-miRNAs) with stem-loop structure. At present, development of computational approach for pre-miRNA identification continues to be a challenging task, in which feature selection is greatly important. Here, we first extracted feature subsets by a hybrid algorithm of genetic algorithm (GA) and support vector machine (SVM) from 124 sequence and secondary structure features. Next, based on the high-frequency features taken from the feature subsets, we proposed a novel stepwise SVM method to identify the optimal feature combinations. The cooperative effect was found among different features in our study. Finally, we obtained 10 feature combinations with strong combined effect which possessed high classification performance for predicting pre-miRNAs. In external validation, all the 10 combinations could predict accurately over 13 pre-miRNAs from 16 new confirmed human pre-miRNAs in miRBase 14.0. The best one could reach 15 (93.75%), which significantly outperformed triplet-SVM (13, 81.25%) in predicting pre-miRNAs.

Key words: MicroRNA precursor, feature selection, genetic algorithm, support vector machine.

INTRODUCTION

MicroRNAs (miRNAs) are a class of non-coding RNAs with size of 21 to 23 nt that are produced from miRNA precursors (pre-miRNAs) of 70 to 90 nt with stem-loop structure by the processing of dicer enzyme (Lim et al., 2003; Bartel, 2004). MiRNAs have been proved to play an important role in post-transcriptional gene regulation in different parts of organisms and at different developmental stages (Lee et al., 2003; Nilsen, 2007), so it is greatly significant to detect miRNAs in various species and further to reveal their function. At present, there are mainly two kinds of methods applied in miRNA prediction: experimental techniques and computational methods. By the experimental method, we only make a portion of high abundance miRNAs cloned effectively, while a large number of low abundance miRNAs and tissue-specific miRNAs are difficult to detect (Bartel, 2004; Wang et al., 2005).

Therefore, in recent years, more and more studies tend to use computational biology methods to identify pre-miRNAs. So far, many algorithms and software based on comparative genomics have been developed, such as miRscan, miRseeker and miRAlign (Lai et al., 2003; Lim et al., 2003; Wang et al., 2005). These approaches adopt sequence conservation to predict the pre-miRNAs, and hence they are difficult to discover the miRNAs which are less or non-conservative between species. In view of the great limitation of comparative genomics methods in

*Corresponding authors. E-mail: zhangruijie2009@yahoo.com.cn, paulsrrao@yahoo.com.cn. Tel/ Fax: +86 451 86615922.

Abbreviations: miRNAs, MicroRNAs; **pre-miRNAs,** miRNA precursors; **SVM,** support vector machine; **GA,** genetic algorithm.

#These authors contributed equally to this work.

predicting non-homologous new miRNAs, researchers began to use computational approaches, particularly machine learning methods to identify pre-miRNAs (Xue et al., 2005; Huang et al., 2007; Jiang et al., 2007). Xue et al. (2005) presented a support vector machine (SVM)-based classifier called triplet-SVM, which classifies human pre-miRNAs from pseudo hairpins based on 32 sequence-structure triplet features. In computational identification study, feature selection is greatly important. However, the features selected in these existing methods were different and would induce different performance (Saeys et al., 2007).

Therefore, how to effectively select a feature subset for pre-miRNA prediction has been always under discussion. Thus we considered scientifically extracting a set of best features to identify pre-miRNAs.

In this study, we obtained 124 sequence and secondary structure features, and developed an algorithm based on R platform (Team, 2009) to perform pre-miRNA feature selection. We used genetic algorithm (GA) to optimize the features and support vector machine (SVM) (Cho and Hermsmeier, 2002; Li et al., 2005; Ng and Mishra, 2007) to classify the two-class samples of real and pseudo pre-miRNAs. Then we obtained the high-frequency features from the feature subsets with high classification accuracy for a further detailed analysis. Since the features are not independent and they have combined effect, we proposed a stepwise SVM algorithm to mine the optimal feature combinations with combined effect from the high-frequency features.

The results showed that the optimal feature combinations we dug out possessed high classification contribution and the prediction accuracy of the best one reached 93.75% in cross validation.

MATERIALS AND METHODS

Datasets

The human pre-miRNAs in our experiment (the positive dataset) were downloaded from miRBase release 13.0 (Griffiths-Jones et al., 2008), which contains 706 reported pre-miRNAs entries from *Homo sapiens*. The negative samples were obtained from the results of Xue et al. (2005), which contains 8494 pretreated non-miRNA hairpin sequences. To balance the sample size, we randomly selected 706 as the negative dataset. In addition, we used five-fold cross validation (*5-fold-cv*) approach to construct the training and testing set; we randomly divided the positive and negative datasets into five non-overlapping subsets of roughly equal size, respectively. A combination of one positive subset and one negative subset constitutes a testing set and the total remaining subsets are used as the training set (Li et al., 2005). Thus the *5-fold-cv* could construct 25 pairs of training and testing sets.

Feature set

We took account of the sequence features and secondary structure features of pre-miRNAs, such as base content, triplet structure, helix structure, loop structure and minimum free energy. The total of 124 features can be divided into six categories (all the features are shown in Supplemental Table 1):

(1) Content of one-dimensional code word (ACGU nucleotide): features F1 - F4;
(2) Content of two-dimensional code word (dinucleotide): features F5 - F20;
(3) Content of three-dimensional code word (trinucleotide): features F21 - F84;
(4) Triplet features (Xue et al., 2005), combining sequence and structural information, describing the matched condition of the three consecutive base pairs and the type of intermediate nucleotide in the sequence: features F85 - F116;
(5) Secondary structure features: features F117 - F123, including "Bulge loop", "External loop", "Hairpin loop", "Helix", "Interior loop", "Multi-loop" and "Stack";
(6) Minimum free energy (MFE): feature F124.

We calculated the values of each feature for the training and testing samples with Perl program.

Experimental methods

We hybridized GA and SVM to obtain a feature subset which can achieve a better classification performance as shown in the left part of Figure 1. We repeated the process of this experiment 100 rounds independently (re-sampling and re-running the program). From the statistics and analysis of these feature subsets in 100 rounds, we obtained the Top20 high-frequency features, and then proposed a novel stepwise SVM method to mine the optimal feature combination which has powerful combined effect and possesses high classification contribution from the Top20 features as shown in the right part of Figure 1.

Feature selection by GA-SVM

Based on the feature values of the sequences, we derived a feature matrix, and then used GA-SVM to perform feature selection. GA is an adaptive global optimization algorithm that simulates biological heredity and evolution in nature. We used SVM classifier with radial basis function (RBF, $k(x, y) = exp(-\gamma \mid x - y \mid^2)$) as kernel function to classify the two types of samples, and then set SVM classification error rate as GA fitness, which is an individual's evaluation indicator of population optimization. The different values of important parameters in GA have a great influence on the results, so we performed the analysis of variance (ANOVA) to select the optimal values of the three primary parameters including genetic generation, population size (the number of individuals in population) and mutation rate. Finally, we selected 100 generations, 50 individuals in the population and mutation rate of 0.05 (Supplemental Table 2 and 3). We randomly generated 50 individuals (*y(i), i = 1, …, 50*), to construct the initial population. $y(i) = (x_1^i, x_2^i, \cdots, x_{124}^i),\ i = 1, \cdots, 50$, where

$$x_j^i = \begin{cases} 1 & \text{if the } j\text{th feature is selected in the individual } i \\ 0 & \text{otherwise} \end{cases},\ i = 1, \cdots, 50,\ j = 1, \cdots, 124$$

We used SVM classifier's error rate as GA fitness, which is used for individual evaluation criteria. We selected half of the population with lower fitness (lower classification error rate) into the next generation directly, then made the remaining half to generate a set of new individuals by crossover and mutation operations with a certain probability. The two parts constituted a new generation population.

We performed this program 100 generations, and then export the

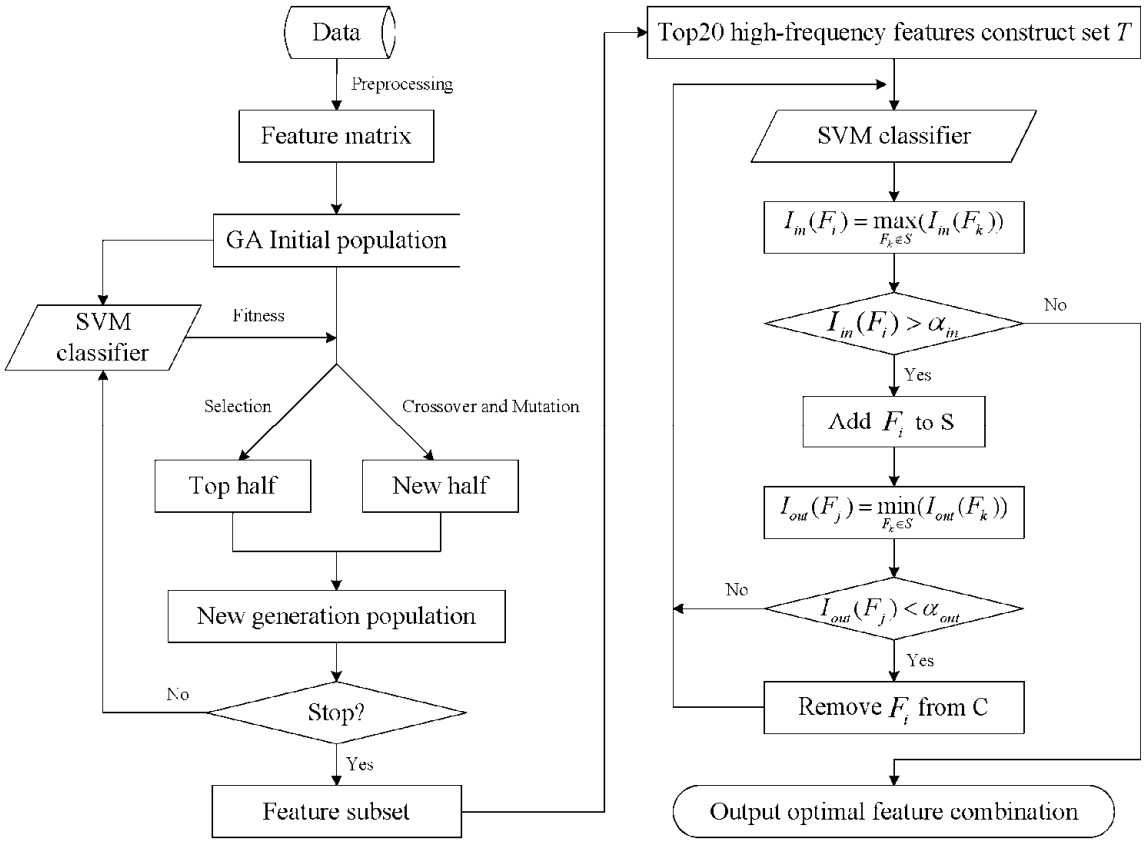

Figure 1. The workflow chart of integrating GA-SVM with stepwise SVM method for predicting pre-miRNAs.

best individual of the last population (see the workflow in Figure 1). Moreover, from the results, we found that there are more than one individuals attaining minimal fitness in the last population; that is, there are many best individuals being able to achieve the minimal classification error rate. Therefore, we extracted a best individual set, including all the best individuals with the minimal fitness from the last population. We supposed that there are *N* best individuals in the best individual set and define a vector $F_{adjusted} = (F(1)/N, F(2)/N, \cdots, F(124)/N)'$, where $F(i),\ i = 1, 2, \cdots 124$, reflects the times of the feature *i* selected in the *N* individuals. Then we set this frequency vector $F_{adjusted}$ as the adjusted best individual which would reflect the relative importance of features in a certain extent.

We repeated the experiments 100 rounds independently, and then we calculated the frequency of each feature that appeared in the adjusted best individual and accumulated these 100 result.

From the result of each round, we found there are almost 20 features in the last generation. So we took out the highest 20 features by the accumulated frequency which we call Top20 high-frequency features to a further detailed analysis.

Stepwise SVM method to mine optimal feature combination

To mine the optimal feature combination, we used the Top20 features forementioned as candidates and used stepwise method to insert or delete a feature at each step.

Step1: We defined two discriminate functions $I_{in}(F_i) = ACC\ (S + F_i) - ACC\ (S)$ and $I_{out}(F_j) = ACC(S) - ACC(S - F_j)$, where *S* is a set of features in the model from previous step, *ACC* ( ) represents the classification accuracy of SVM classifier with features in the brackets, and two threshold values α_{in} and α_{out}. Here, we set α_{in} and $\alpha_{out} = 0.001$.

Step 2: Calculate $ACC(F_i)$, $i = 1, 2, \ldots, 20$, respectively. If ACC (F_i)= max (ACC(F_k)), we would firstly add feature F_i into the training model.

Step 3: Suppose there are $m(1 < m \le 20)$ features in the model, we mark these features as a set *S*, and then calculate $I_{in}(F_i)$ for each F_i not in *S*. If $I_{in}(F_i) = \max_{F_k \in S}(I_{in}(F_k))$ and $I_{in}(F_i) > \alpha_{in}$, we would add the feature F_i into the training model and turn to step4, otherwise turn to step5.

Step 4: Suppose there are $n(1 < n \le 20)$ features in the model, we mark these features as a set *S* and calculate $I_{out}(F_j)$ for each F_j in *S*. If $I_{out}(F_j) = \min_{F_k \in S}(I_{out}(F_k))$ and $I_{out}(F_j) < \alpha_{out}$, we would remove the feature F_j from the training model. And then turn to step3.

Step 5: We obtained a feature combination without adding or removing any features. Then we do *5-fold-cv* to evaluate prediction performance. All the programs are developed on R platform.

Prediction performance assessment

For a prediction problem, a classifier can classify an individual instance into the following four categories: true positive (*TP*), false positive (*FP*), true negative (*TN*) and false negative (*FN*). The specificity (*SP*), sensitivity (*SE*) and total prediction accuracy (*ACC*) are the popular indices for assessment of the prediction system (Jiang et al., 2007; Hand, 2009). They are given by:

$$SP = TN/(TN + FP) \times 100\%$$

$$SE = TP/(TP + FN) \times 100\%$$

$$ACC = (TP + TN)/(TP + FN + TN + FP) \times 100\%$$

RESULTS AND DISCUSSION

Feature subsets extracted by GA-SVM method

With the optimization of each generation, the classification accuracy increases gradually as shown in Figure 2. Within about 20 generations, the changing speed of *ACC* is very rapid, after that it changes slowly and fluctuates slightly in a lesser extent, but the general trend still gradually rises. Further, to take account of the stability of the feature subsets, we repeated the procedure 100 rounds independently (re-sampling and running the program). As a result, the feature subsets in different rounds were different. From each round, we could get an adjusted best individual. Then we accumulated these 100 results to get the frequency diagram of each feature. The results are shown in Figure 3. We sorted the features in descending order of the frequency and took out Top20 high-frequency features as shown in Figure 4. These Top20 features cover all the six categories (Table 1).

Optimal feature combination mined by stepwise SVM method

We used stepwise SVM method to mine the optimal feature combination. The whole process was repeated 10 times and the results are shown in Table 2. As shown in Table 2, we found that the optimal combinations extracted from these 10 rounds were variously different, but we discovered some commonness of these combinations.

Feature F124 (MFE) and F3 (G content) were extremely stable, appearing in all the rounds, and their ranks are quite high. This is consistent with previous correlative conclusion. The feature F124 is MFE which is ranked first of the Top20 high-frequency features. It appears in each optimal feature subset and has strong classification ability. A large number of studies also point out that MFE is an important feature in the distinction between pre-miRNAs and ordinary hairpin sequences (Hofacker, 2003; Jiang et al., 2007). The feature F3 is the content of guanine (G) which is one of DNA's four bases, possessing the smallest adiabatic ionization potential of the four, so the oxidation of DNA usually occurs in Guanine. The guanine and cytosine content may influence the thermodynamic stability of a pre-miRNA molecule (Koparde et al., 2010).

Feature F100 (C+++) or F2 (C content) will certainly appears, but generally the two do not appear at the same time (only simultaneously appearing in the fifth round). F100 is the structure feature C+++, which describes the number of triplet structure that three consecutive base pairs are fully matched and the middle base is cytosine. It is ranked 15th in Xue et al.'s (2005) and 18th in Jiang et al.'s (2007) reports.

However, in our study it was ranked first in the 32 triplet features (6th in the Top20 features). Xue et al. (2005) used only these 32 triplet features and Jiang et al. (2007) used two more features about MFE, but we used much more other types of features. We know that there are certain influence and interaction between the features, so our results are much more believable and demonstrate that the C+++ is a crucial feature and plays an important role in the classification.

Feature F123 (Stack statistic) will generally be the first selected and rarely removed in the end (only be removed once in the eighth round). Feature F123 is stack, and it ranked 8th in the Top20 features. The two contiguous base pairs stacking together formed a stack. Nucleic acids are stabilized by base stacking (Gabb et al., 1996). The various loops and the stacks constitute the secondary structure of pre-miRNAs. The energy of the secondary structure is assumed to be the sum of the energy contributions of loops and stacks (Hofacker et al., 1994, 2004). The base stacking could improve the thermal stability (Chen et al., 2005).

The frequency of trinucleotide feature F41 (GCA content) is extremely high. There are 7 combinations (a total of 10) including this feature. The phenomenon was also found in previous experiments as shown in Figure 4. In particular, we noticed that GCA content was ranked third in front of GC content (the best in dinucleotide features) which is ranked 7th. Thus, it can be seen that this feature has a very strong ability in classification.

Although, we did not get a unique feature combination, we discovered some regularity among them. Each combination has almost included these six classes of features and just needed to select one or two special features from each class to identify pre-miRNAs and can obtain a very good classification result. In particular, the smallest combination only contains 5 features from four categories. Such discovery is significant to research the characters of miRNAs. We can get a classification accuracy of 92.92% using only 6 features in our best combination, 92.07% using the Top20 features and 89.01% using all the 124 features (Figure 5, Table 2). We obtained the higher accuracy with the fewer features by our stepwise SVM method, thus indicating that it was powerful.

Table 1. The summary of Top20 features belonging to six categories.

Feature category	Included / Total	Included features list
Nucleotide	2 / 4	G, C
Dinucleotide	6 / 16	GC, UA, GG, UC, GU, GA
Trinucleotide	5 / 64	GCA, UUA, AAC, GGC, UAA
Triplet	4 / 32	C+++, U+++, A+++, G++
Secondary structure	2 / 7	Stack, Interior-loop
Minimum free energy	1 / 1	MFE

Table 2. The combinations mined from Top20 features.

Rank	Feature combinations (listed from left to right by the adding order)											*ACC* (%)
1	F123	F3	F124	F2	F16	F41						92.92
2	F123	F3	F124	F2	F16	F116	F81	F47	F105	F41	F121	92.81
3	F123	F100	F3	F124	F16	F41	F47	F14				92.51
4	F123	F2	F124	F3	F41	F105	F23					92.47
5	F123	F2	F124	F3	F16	F116	F100	F105				92.44
6	F123	F2	F124	F3	F121	F41						92.13
7	F123	F2	F124	F3	F14							92.13
8	F100	F3	F124	F16	F13	F81	F47	F41				91.84
9	F123	F100	F3	F124	F41	F17						91.73
10	F123	F100	F3	F124	F81							90.94
Average												**92.19**

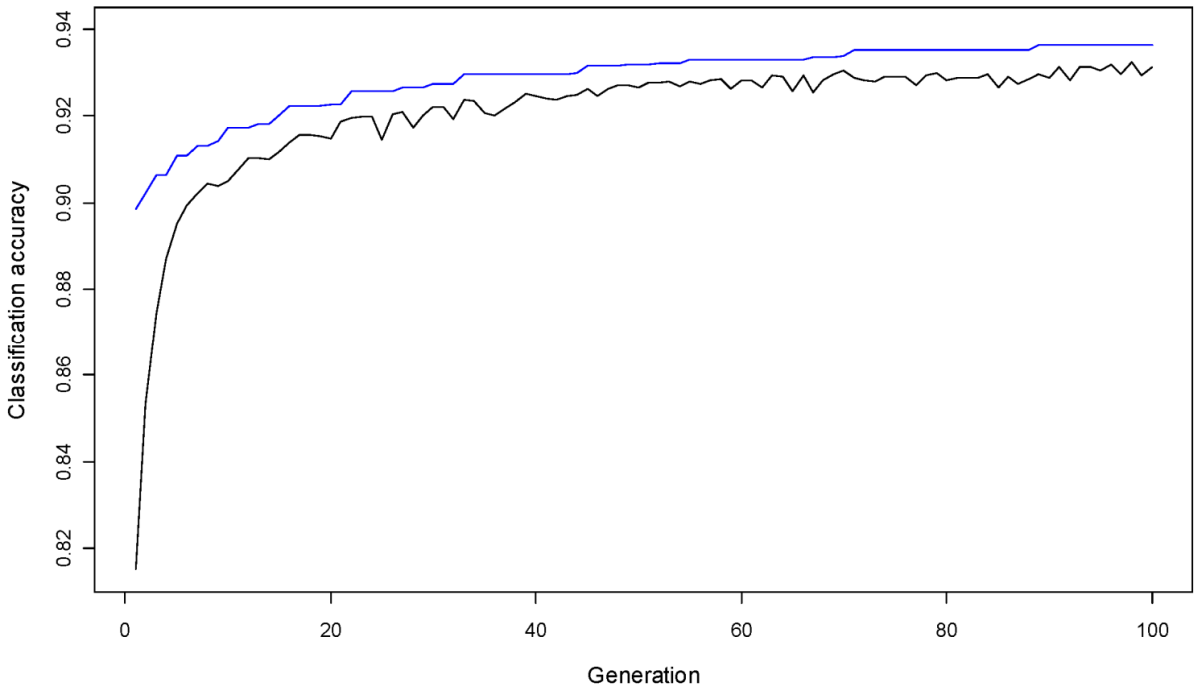

Figure 2. The changing curve of the best individual's classification accuracy and the average *ACC* of the whole population in each generation. The blue line represents the changing trend of the best individual's classification accuracy in each generation, and the black line stands for the average *ACC* of all individuals in each generation population.

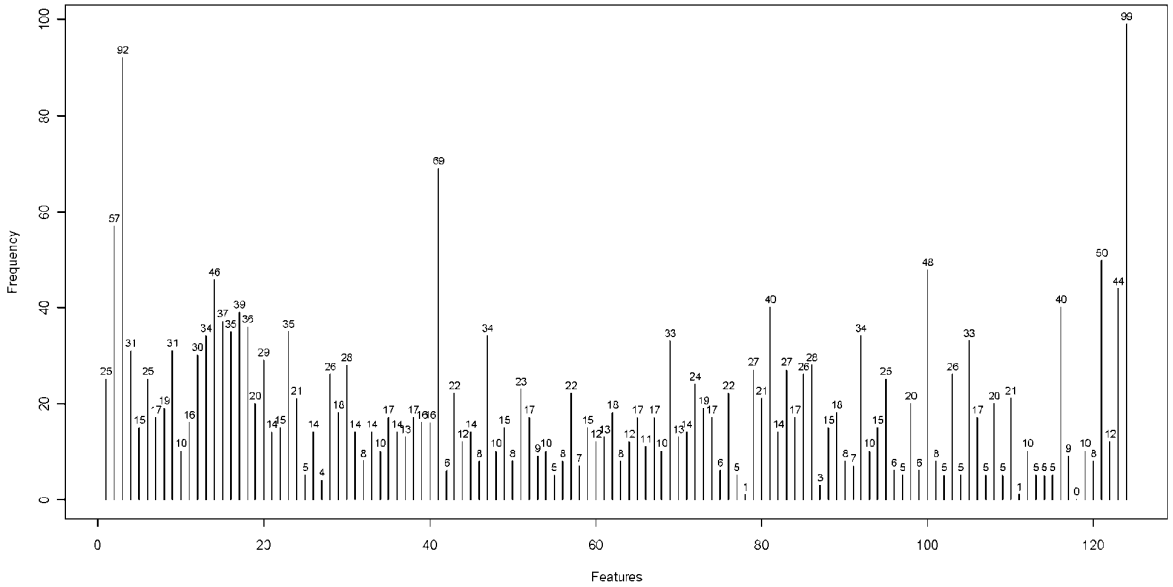

Figure 3. The cumulative frequency graph of the feature subsets in 100 rounds.

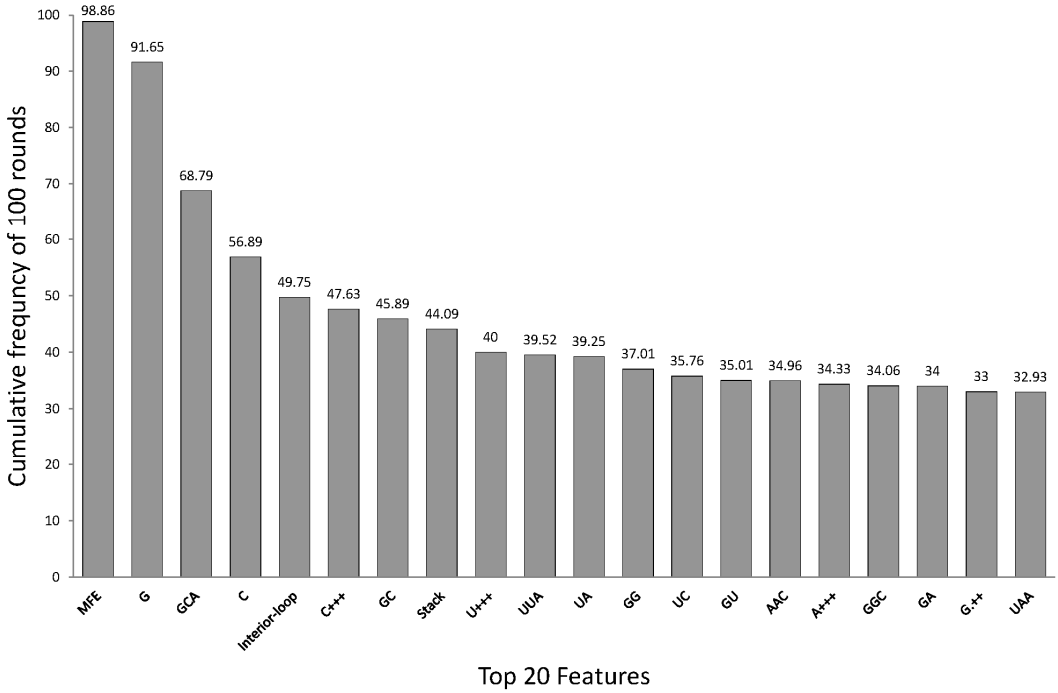

Figure 4. Top20 high-frequency features.

Performance evaluation of optimal feature combinations using *5-fold-cv*

To evaluate the classification performance of the identified optimized feature combinations, we compared our method to triplet-SVM (Xue et al., 2005). First, we used 706 human pre-miRNAs in miRBase13.0 as the positive set, and randomly selected 706 from 8494 negative samples as the negative set. Next, *5-fold-cv* based on SVM was performed on our 10 feature combinations and the feature set of triplet-SVM. Finally, the classification performance was evaluated by three indices: *SP*, *SE* and *ACC*. As a result, our 10 optimal feature combinations averagely achieved the higher *SP*

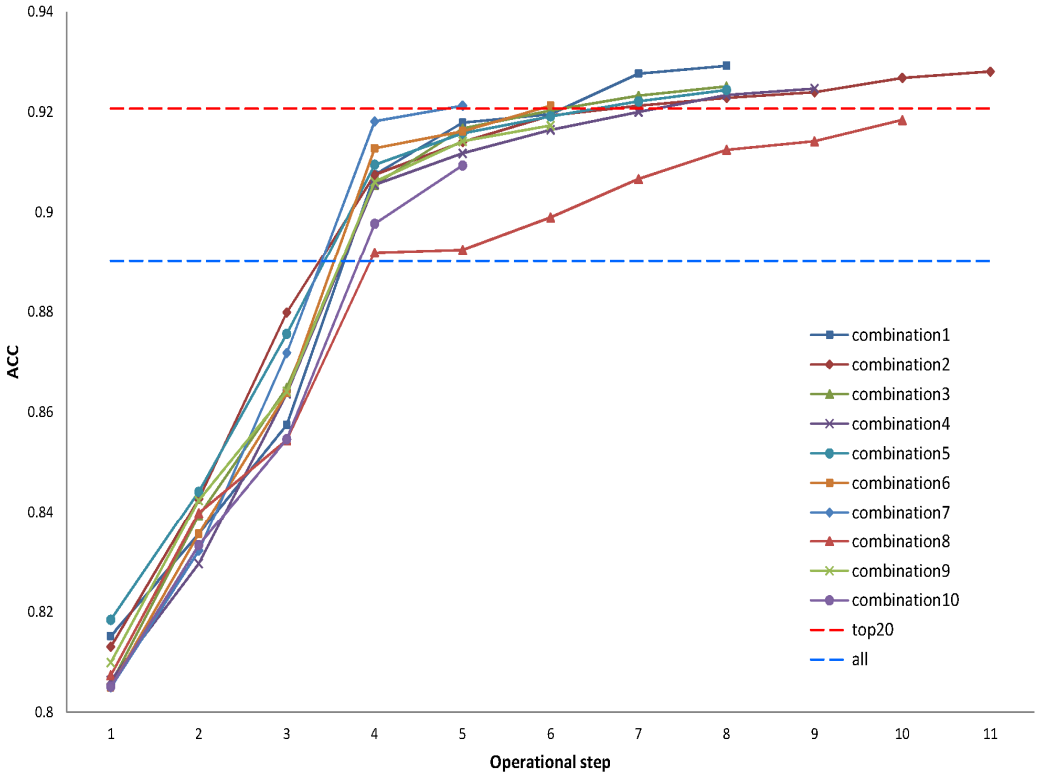

Figure 5. The changing of *ACC* in stepwise SVM method. Each line represents a feature combination. The red dotted line (*ACC*=92.07%) represents the classification accuracy using the Top20 high-frequency features and the blue dotted line (*ACC*=89.01%) represents the classification accuracy using all the 124 features.

Table 3. Performance evaluation of feature combinations using *5-fold-cv*.

Feature combinations	*SP* (%)	*SE* (%)	*ACC* (%)
Combination 1	**96.34**	**89.50**	**92.92**
Combination 2	95.43	89.93	92.68
Combination 3	95.49	89.70	92.60
Combination 4	95.35	89.02	92.18
Combination 5	95.60	89.30	92.45
Combination 6	94.92	89.28	92.10
Combination 7	94.72	89.48	92.10
Combination 8	95.80	88.82	92.31
Combination 9	94.27	89.08	91.67
Combination 10	93.70	88.94	91.32
Average	**95.16**	**89.30**	**92.23**
Triplet-SVM	**91.86**	**83.15**	**87.50**

(95.16%), *SE* (89.30%) and *ACC* (92.23%) than that of triplet-SVM (91.86%, 83.15% and 87.50%, seen in Table 3). Especially, the best one achieved 96.34% (*SP*), 89.50 (*SE*) and 92.92% (*ACC*) as shown in Table 3. Thus, the optimal feature combinations were effective for distinguishing real pre-miRNAs from pseudo pre-miRNAs.

Validation by the latest confirmed human pre-miRNAs

To test the predicted performance of the feature combinations extracted, we used the optimal feature combinations to predict the 16 new human pre-miRNAs

Table 4. The predicted results using the optimal feature combinations and triplet-SVM.

Feature combinations		16 new pre-miRNAs (miRBase14.0)	
Combination index	**Number of features**	**Number of correctly identified**	**Predict accuracy (%)**
Combination 1	6	14	87.50
Combination 2	**11**	**15**	**93.75**
Combination 3	8	13	81.25
Combination 4	7	14	87.50
Combination 5	8	13	81.25
Combination 6	6	14	87.50
Combination 7	5	13	81.25
Combination 8	8	14	87.50
Combination 9	6	14	87.50
Combination 10	5	14	87.50
Triplet-SVM	**32**	**13**	**81.25**

confirmed in miRBase 14.0. First, the 706 pre-miRNAs in miRBase13.0 and the re-extracted 706 negative samples constituted the training set. Then the classifier model based on our optimal feature combinations was built to predict the 16 new confirmed human pre-miRNAs. Next, we performed triplet-SVM (Xue et al., 2005) to predict the 16 pre-miRNAs. The results are shown in Table 4. The number of correctly identified of each combination can reach over 13, wherein six combinations reach 14. The best combination correctly identified 15 (93.75%) with 11 features and triplet-SVM correctly identified 13 (81.25%) with 32 features. Obviously, compared with triplet-SVM, we obtained higher validation result based on fewer features.

Conclusion

In this study, we focused on the optimization of feature combination in predicting pre-miRNAs. We performed the feature selection applying GA-SVM method and further extracted the feature combination with strong combined effect and high classification performance applying a novel stepwise SVM method that proved to be of great significance. The results show that the features MFE, G content and GCA content are crucial to distinguish pre-miRNAs. Although, we did not find the only one optimal feature combination, we discovered some regularity among them. Each combination has almost included these six categories of features, and we just needed to select one or two special features from each category to identify pre-miRNAs. Furthermore, we discovered an important feature “stack” that did not get enough attention before. Its classification performance was very strong which could be almost the first selected in the stepwise mining procedure.

Compared to the previous study on triplet-SVM, our optimal feature combinations based on our stepwise SVM method were ~5% higher in sensitivity, specificity and accuracy by *5-fold-cv*. Through the prediction of 16 new validated pre-miRNAs in miRBase 14.0, we also correctly identified significantly more pre-miRNAs than triplet-SVM (15 (93.75%) with the best combination vs. 13 (81.25%) with triplet-SVM). We believe that our stepwise SVM method effectively reduced the number of features and improved the classification performance. In addition, the optimal feature combinations we mined are effective for miRNAs prediction and characteristic research.

Supplementary Material

Supplemental table 1 is the list of all the 124 features, and supplemental table 2 and 3 represent the result of GA parameter optimization by ANOVA. Available at http://www.academicjournals.org/AJB.

ACKNOWLEDGEMENTS

This work was partly supported by the the National Natural Science Foundation of China (Grant no. 81172842), National Science Foundation of Heilongjiang Province (Grant no. F2008-02) and the Excellent Youth Foundation of Heilongjiang Province (Grant no. JC200711).

REFERENCES

Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116(2): 281-297.

Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ (2005). Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33(20): e179.

Cho SJ, Hermsmeier MA (2002). Genetic Algorithm guided Selection: variable selection and subset selection. J. Chem. Inf. Comput. Sci. 42(4): 927-936.

Gabb HA, Sanghani SR, Robert CH, Prevost C (1996). Finding and

visualizing nucleic acid base stacking. J. Mol. Graph, 14(1): 6-11, 23-14.
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res. 36(Database issue): D154-158.
Hand DJ (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn, 77: 103-123.
Hofacker IL (2003). Vienna RNA secondary structure server. Nucleic Acids Res. 31(13): 3429-3431.
Hofacker IL (2004). RNA secondary structure analysis using the Vienna RNA package. Curr. Protoc. Bioinformatics, Chapter 12: Unit 12 p. 12.
Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994). Fast Folding and Comparison of RNA Secondary Structures. Monatshefte fuer Chemie, 125: 167-188.
Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, Zhao SH (2007). MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics, 8: p. 341.
Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 35(Web Server issue): W339-344.
Koparde P, Singh S (2010). Prediction of micro RNAs against H5N1 and H1N1 NS1 Protein: a Window to Sequence Specific Therapeutic Development. J Data Mining in Genom Proteomics, 1(2): p. 104.
identification of Drosophila microRNA genes. Genome Biol. 4(7): R42.
Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature, 425(6956): 415-419.
Li L, Jiang W, Li X, Moser KL, Guo Z, Du L, Wang Q, Topol EJ, Rao S (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 85(1): 16-23.
Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP (2003). The microRNAs of Caenorhabditis elegans. Genes Dev. 17(8): 991-1008.
Ng KL, Mishra SK (2007). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 23(11): 1321-1330.
Nilsen TW (2007). Mechanisms of microRNA-mediated gene regulation in animal cells. Trends Genet. 23(5): 243-249.
Saeys Y, Inza I, Larranaga P (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19): 2507-2517.
Team RDC (2009). R: A Language and Environment for Statistical Computing. Available from http://www.R-project.org.
Wang X, Zhang J, Li F, Gu J, He T, Zhang X, Li Y (2005). MicroRNA identification based on sequence and structure alignment. Bioinformatics 21(18): 3610-3614.
Xue C, Li F, He T, Liu GP, Li Y, Zhang X (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics, 6: p. 310.

Supplemental table 1. Feature list.

Feature number	Feature	Feature categories
F1	A	Nucleotide (content of one-dimensional code word)
F2	C	Nucleotide (content of one-dimensional code word)
F3	G	Nucleotide (content of one-dimensional code word)
F4	U	Nucleotide (content of one-dimensional code word)
F5	AA	Dinucleotide (content of two-dimensional code word)
F6	AC	Dinucleotide (content of two-dimensional code word)
F7	AG	Dinucleotide (content of two-dimensional code word)
F8	AU	Dinucleotide (content of two-dimensional code word)
F9	CA	Dinucleotide (content of two-dimensional code word)
F10	CC	Dinucleotide (content of two-dimensional code word)
F11	CG	Dinucleotide (content of two-dimensional code word)
F12	CU	Dinucleotide (content of two-dimensional code word)
F13	GA	Dinucleotide (content of two-dimensional code word)
F14	GC	Dinucleotide (content of two-dimensional code word)
F15	GG	Dinucleotide (content of two-dimensional code word)
F16	GU	Dinucleotide (content of two-dimensional code word)
F17	UA	Dinucleotide (content of two-dimensional code word)
F18	UC	Dinucleotide (content of two-dimensional code word)
F19	UG	Dinucleotide (content of two-dimensional code word)
F20	UU	Dinucleotide (content of two-dimensional code word)
F21	AAA	Trinucleotide (content of three-dimensional code word)

Supplemental table 1. Contd

F22	AAG	Trinucleotide (content of three-dimensional code word)
F23	AAC	Trinucleotide (content of three-dimensional code word)
F24	AAU	Trinucleotide (content of three-dimensional code word)
F25	ACA	Trinucleotide (content of three-dimensional code word)
F26	ACG	Trinucleotide (content of three-dimensional code word)
F27	ACC	Trinucleotide (content of three-dimensional code word)
F28	ACU	Trinucleotide (content of three-dimensional code word)
F29	AGA	Trinucleotide (content of three-dimensional code word)
F30	AGG	Trinucleotide (content of three-dimensional code word)
F31	AGC	Trinucleotide (content of three-dimensional code word)
F32	AGU	Trinucleotide (content of three-dimensional code word)
F33	AUA	Trinucleotide (content of three-dimensional code word)
F34	AUG	Trinucleotide (content of three-dimensional code word)
F35	AUC	Trinucleotide (content of three-dimensional code word)
F36	AUU	Trinucleotide (content of three-dimensional code word)
F37	GAA	Trinucleotide (content of three-dimensional code word)
F38	GAG	Trinucleotide (content of three-dimensional code word)
F39	GAC	Trinucleotide (content of three-dimensional code word)
F40	GAU	Trinucleotide (content of three-dimensional code word)
F41	GCA	Trinucleotide (content of three-dimensional code word)
F42	GCG	Trinucleotide (content of three-dimensional code word)
F43	GCC	Trinucleotide (content of three-dimensional code word)
F44	GCU	Trinucleotide (content of three-dimensional code word)
F45	GGA	Trinucleotide (content of three-dimensional code word)
F46	GGG	Trinucleotide (content of three-dimensional code word)
F47	GGC	Trinucleotide (content of three-dimensional code word)
F48	GGU	Trinucleotide (content of three-dimensional code word)
F49	GUA	Trinucleotide (content of three-dimensional code word)
F50	GUG	Trinucleotide (content of three-dimensional code word)
F51	GUC	Trinucleotide (content of three-dimensional code word)

Supplemental table 1. Contd.

F52	GUU	Trinucleotide (content of three-dimensional code word)
F53	CAA	Trinucleotide (content of three-dimensional code word)
F54	CAG	Trinucleotide (content of three-dimensional code word)
F55	CAC	Trinucleotide (content of three-dimensional code word)
F56	CAU	Trinucleotide (content of three-dimensional code word)
F57	CCA	Trinucleotide (content of three-dimensional code word)
F58	CCG	Trinucleotide (content of three-dimensional code word)
F59	CCC	Trinucleotide (content of three-dimensional code word)
F60	CCU	Trinucleotide (content of three-dimensional code word)
F61	CGA	Trinucleotide (content of three-dimensional code word)
F62	CGG	Trinucleotide (content of three-dimensional code word)
F63	CGC	Trinucleotide (content of three-dimensional code word)
F64	CGU	Trinucleotide (content of three-dimensional code word)
F65	CUA	Trinucleotide (content of three-dimensional code word)
F66	CUG	Trinucleotide (content of three-dimensional code word)
F67	CUC	Trinucleotide (content of three-dimensional code word)
F68	CUU	Trinucleotide (content of three-dimensional code word)
F69	UAA	Trinucleotide (content of three-dimensional code word)

Supplemental table 1. Contd.

F70	UAG	Trinucleotide (content of three-dimensional code word)
F71	UAC	Trinucleotide (content of three-dimensional code word)
F72	UAU	Trinucleotide (content of three-dimensional code word)
F73	UCA	Trinucleotide (content of three-dimensional code word)
F74	UCG	Trinucleotide (content of three-dimensional code word)
F75	UCC	Trinucleotide (content of three-dimensional code word)
F76	UCU	Trinucleotide (content of three-dimensional code word)
F77	UGA	Trinucleotide (content of three-dimensional code word)
F78	UGG	Trinucleotide (content of three-dimensional code word)
F79	UGC	Trinucleotide (content of three-dimensional code word)
F80	UGU	Trinucleotide (content of three-dimensional code word)
F81	UUA	Trinucleotide (content of three-dimensional code word)
F82	UUG	Trinucleotide (content of three-dimensional code word)
F83	UUC	Trinucleotide (content of three-dimensional code word)
F84	UUU	Trinucleotide (content of three-dimensional code word)
F85	A...	Triplet feature
F86	A..+	Triplet feature
F87	A.+.	Triplet feature
F88	A+..	Triplet feature
F89	A.++	Triplet feature
F90	A+.+	Triplet feature
F91	A++.	Triplet feature
F92	A+++	Triplet feature
F93	C...	Triplet feature
F94	C..+	Triplet feature
F95	C.+.	Triplet feature
F96	C+..	Triplet feature
F97	C.++	Triplet feature
F98	C+.+	Triplet feature
F99	C++.	Triplet feature
F100	C+++	Triplet feature
F101	G...	Triplet feature
F102	G..+	Triplet feature
F103	G.+.	Triplet feature

Supplemental table 1. Contd.

F104	G+..	Triplet feature
F105	G.++	Triplet feature
F106	G+.+	Triplet feature
F107	G++.	Triplet feature
F108	G+++	Triplet feature
F109	U...	Triplet feature
F110	U..+	Triplet feature
F111	U.+.	Triplet feature
F112	U+..	Triplet feature
F113	U.++	Triplet feature
F114	U+.+	Triplet feature
F115	U++.	Triplet feature
F116	U+++	Triplet feature
F117	Bulge-loop	Secondary structural feature

Supplemental table 1. Contd.

F118	External-loop	Secondary structural feature
F119	Hairpin-loop	Secondary structural feature
F120	Helix	Secondary structural feature
F121	Interior-loop	Secondary structural feature
F122	Multi-loop	Secondary structural feature
F123	Stack	Secondary structural feature
F124	MFE	Minimal folding free energy

Supplemental table 2. Analysis of variance for different parameters in GA.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.000	11	2.77E-005	1.613	.125
Intercept	.264	1	.264	15325.367	.000
Genetic algebra	2.25E-005	1	2.25E-005	1.308	.258
Population size	5.75E-006	1	5.75E-006	.335	.566
Mutation rate	.000	2	.000	7.830	.001 (*)
Genetic generation * Population size	3.03E-006	1	3.03E-006	.176	.677
Genetic generation * Mutation rate	1.66E-006	2	8.29E-007	.048	.953
Population size * Mutation rate	2.41E-006	2	1.20E-006	.070	.933
Genetic generation * Population size * Mutation rate	4.55E-007	2	2.28E-007	.013	.987
Error	.001	48	1.72E-005		
Total	.265	60			
Corrected Total	.001	59			

Dependent Variable: Classification error rate
*The mean difference is significant at the .05 level.

Supplemental table 3. Multiple Comparison for mutation rate.

	(I) Mutation rate	(J) Mutation rate	Mean Difference (I-J)	Std. Error	Sig.
Tukey HSD	1.00	2.00	-.0047 (*)	.00131	.002
		3.00	-.0043 (*)	.00131	.006
	2.00	1.00	.0047 (*)	.00131	.002
		3.00	.0004	.00131	.942
	3.00	1.00	.0043 (*)	.00131	.006
		2.00	-.0004	.00131	.942
LSD	1.00	2.00	-.0047 (*)	.00131	.001
		3.00	-.0043 (*)	.00131	.002
	2.00	1.00	.0047 (*)	.00131	.001
		3.00	.0004	.00131	.743
	3.00	1.00	.0043 (*)	.00131	.002
		2.00	-.0004	.00131	.743

Dependent Variable: Classification error rate
Based on observed means.
*The mean difference is significant at the .05 level.