Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

# Effective sample selection for classification of pre-miRNAs

**K. Han**

School of Computer and Information Engineering,
Harbin University of Commerce, Harbin, Heilongjiang, China

Corresponding author: K. Han
E-mail: hanke@hrbcu.edu.cn

Genet. Mol. Res. 10 (1): 506-518 (2011)
Received September 8, 2010
Accepted October 26, 2010
Published March 29, 2011
DOI 10.4238/vol10-1gmr1054

**ABSTRACT.** To solve the class imbalance problem in the classification of pre-miRNAs with the *ab initio* method, we developed a novel sample selection method according to the characteristics of pre-miRNAs. Real/pseudo pre-miRNAs are clustered based on their stem similarity and their distribution in high dimensional sample space, respectively. The training samples are selected according to the sample density of each cluster. Experimental results are validated by the cross-validation and other testing datasets composed of human real/pseudo pre-miRNAs. When compared with the previous method, *microPred*, our classifier *miRNAPred* is nearly 12% more accurate. The selected training samples also could be used to train other SVM classifiers, such as *triplet-SVM*, *MiPred*, *miPred*, and *microPred*, to improve their classification performance. The sample selection algorithm is useful for constructing a more efficient classifier for the classification of real pre-miRNAs and pseudo hairpin sequences.

**Key words:** Sample selection; Class imbalance; Pre-miRNA; Information gain; Conservation

 

## INTRODUCTION

MicroRNAs (miRNA) are non-coding RNAs about 21 to 24 nucleotides (nt) in length, which can play important roles in gene regulation by targeting mRNAs for cleavage or translational repression (Bartel, 2004; Chatterjee and Grosshans, 2009). It has been shown that miRNAs usually participate in the important life processes, including growth process, hemopoiesis, organ formation, apoptosis, and cell proliferation. Furthermore, they are closely related to many kinds of human diseases including cancer (Bushati and Cohen, 2007). Due to the difficulty of systematically detecting miRNAs from a genome by existing experimental techniques, computational methods play important roles in the identification of new miRNAs.

The 60- to 70-nt precursor miRNAs (pre-miRNAs) have the characteristic of stem-loop hairpin structures, which is an important feature used in the computational identification of miRNAs. Recently, the *ab initio* method based on machine learning was presented to distinguish real pre-miRNAs from candidate hairpin sequences. Real/pseudo pre-miRNAs are selected as positive/negative training samples to construct the classifiers. These classifiers include support vector machines (SVM) (Xue et al., 2005; Ng and Mishra, 2007; Batuwita and Palade, 2009), probabilistic co-learning model (Nam et al., 2005), naive Bayes (Yousef et al., 2006, 2008), random forest (Jiang et al., 2007), and kernel density estimation (Chang et al., 2008). A hairpin-like candidate sequence could be classified as a real pre-miRNA or a pseudo pre-miRNA by the classifiers.

The positive and negative training datasets would greatly affect the classification performance. Recently, only several thousands of miRNAs have been identified and verified by biological experiments. Obviously, known miRNAs are much less in a single species. However, millions of hairpin-like pseudo sequences could be obtained from the genome segment of human. *Triplet-SVM* (Xue et al., 2005) obtained 8494 human pseudo hairpin sequences from the protein coding regions, and there were just 193 known real pre-miRNAs at that time. These 8494 human pseudo hairpin sequences are often used in the classification and prediction of pre-miRNAs with the *ab initio* method. *Micropred* (Batuwita and Palade, 2009) could obtain 691 real pre-miRNAs and 9248 pseudo sequences, which include 8494 pseudo hairpins and 754 other non-coding RNAs (ncRNAs). The ratio of the positive to negative dataset is 1:13.4. Therefore, when classifying real pre-miRNAs and pseudo pre-miRNAs, the main problem encountered in the dataset is its imbalance. The current research has shown that training a classifier with such an imbalance positive and negative dataset could result in poor classification performance with respect to the minority class (Weiss, 2004). It is essential to select appropriate representative sample subsets to train the classifier, which contributes to improving classification accuracy.

It has been found that SVM classifiers can also be sensitive to the class imbalance (Veropoulos et al., 1999; Akbani et al., 2004). Yousef et al. (2008) created a one-class SVM (OC-SVM) that depends only on positive samples (real pre-miRNAs). OC-SVM could eliminate the imbalance problem. However, the classification performance of one-class machine learning is lower than the one of two-class machine learning. The *triplet-SVM*, *MiPred* and *miPred* methods chose real/pseudo pre-miRNAs randomly from the imbalanced dataset as the training samples. These samples could not represent positive and negative sample spaces completely and decreased the accuracy of the classifiers.

 

*MicroPred* evaluated the methods of processing imbalance, including random over/under-sampling (Weiss, 2004), *SMOTE* (Chawla et al., 2002), multi-classifier system training (Molinara et al., 2007), different error costs (Veropoulos et al., 1999; Akbani et al., 2004), and *zSVM* (Imam et al., 2006). The classifier trained with the *SMOTE* method achieved the best performance. However, *SMOTE* assumes that the sample locating between a positive/negative sample and its positive/negative neighbor is also positive/negative. The distribution of real and pseudo pre-miRNAs could not satisfy this assumption well. Moreover, the *SMOTE* method introduces newly generated pre-miRNAs into the positive sample space, which changes the original distribution of positive samples (real pre-miRNAs).

Wang et al. (2010; our group members) solved the class imbalance problem in mining single nucleotide polymorphisms (SNPs) with ensemble classifiers. The pseudo pre-miRNAs are extracted through searching hairpin-like sequences from the genome segment of human. It is difficult to separate the majority of pseudo pre-miRNAs into several subsets. Therefore, the method with ensemble classifiers does not apply to the case about pre-miRNAs.

## MATERIAL AND METHODS

### Characteristics of pre-miRNAs

#### *Features of pre-miRNAs*

Recent research indicates that pre-miRNAs have many features about both primary sequence and secondary structure. These features are typically used to construct a classifier to classify the real pre-miRNAs and pseudo hairpin sequences.

*miPred* (Ng and Mishra, 2007) extracted 29 global and intrinsic folding features from human real/pseudo pre-miRNAs. These features are: 1) Seventeen base composition variables, including 16 dinucleotide frequencies, that is *XY*% where $X,Y \in \{A,C,G,U\}$, and $(G + C)$% content; 2) Six folding measures: adjusted base pairing propensity *dP* (Schultes et al., 1999), adjusted minimum free energy of folding (*MFE*) denoted as *dG* (Seffens and Digby, 1999; Freyhult et al., 2005), adjusted base pair distance *dD* (Moulton et al., 2000; Freyhult et al., 2005), adjusted Shannon entropy *dQ* (Freyhult et al., 2005), MFE index 1, $MFEI_1$ (Zhang et al., 2006), and MFE index 2, $MFEI_2$; 3) One topological descriptor, which is the degree of compactness *dF* (Fera et al., 2004; Gan et al., 2004), and 4) Five normalized variants of *dP*, *dG*, *dQ*, *dD*, and *dF*: *zP*, *zG*, *zQ*, *zD*, and *zF*.

In addition to the above 29 features, *microPred* extracted 19 new features, totaling 48 features. These features are: 1) Two minimum free energy-related features: MFE index 3 ($MFEI_3$) and MFE index 4 ($MFEI_4$); 2) Four RNAfold-related features: normalized ensemble free energy (*NEFE*), frequency of the MFE structure *Freq*, structural diversity denoted as *Diversity*, and a combined feature *Diff*; 3) Six thermodynamic features: structure entropy *dS* and *dS/L*, structure enthalpy *dH* and *dH/L*, melting energy of the structure $T_m$ and $T_m/L$, where *L* is the length of pre-miRNA sequence, and 4) Seven base pair-related features: |*A-U*|/*L*, |*G-C*|/*L*, |*G-U*|/*L*, average base pairs per stem *Avg_BP_Stem*, (*A-U*)*%/n_stems*, (*G-C*)*%/n_stems*, (*G-U*)*%/n_stems*, where *n_stems* is the number of stems in the secondary structure.

The above 48 features could fully reflect the distribution of positive/negative samples (real/pseudo pre-miRNAs) in sample space (Batuwita and Palade, 2009). Therefore, clustering samples should consider the high dimensional characteristic of pre-miRNAs.

### *Conservation of pre-miRNAs*

Pre-miRNAs are typically about 60 to 70 nt, and contain a ~22-bp double-stranded stem and a ~10-nt terminal loop. Recently, computational phylogenetic shadowing showed that the stems of pre-miRNAs are highly conserved in whole genome alignments, whereas most terminal loop sequences are only loosely conserved (Berezikov et al., 2005). Thus, if the stems of two pre-miRNAs are more consistent, these two pre-miRNAs are more similar. The similarity is measured through observing the consistent degree of nucleotide sequence in stems. Therefore, the pre-miRNAs with similar stems should be gathered into the same cluster.

## Sample selection algorithm

According to above the characteristics of pre-miRNAs, we proposed a sample selection algorithm as shown in Figure 1. 1) The samples (real/pseudo pre-miRNAs) are clustered according to their stem similarity. 2) The discriminative feature subset is selected to better reflect the distribution of samples. The feature selection considers the feature difference and information gain. 3) The samples are clustered according to their distribution in sample space. 4) The training samples are selected according to the sample density of each cluster. 5) An SVM classifier is trained with our selected samples to validate the performance of sample selection algorithm.

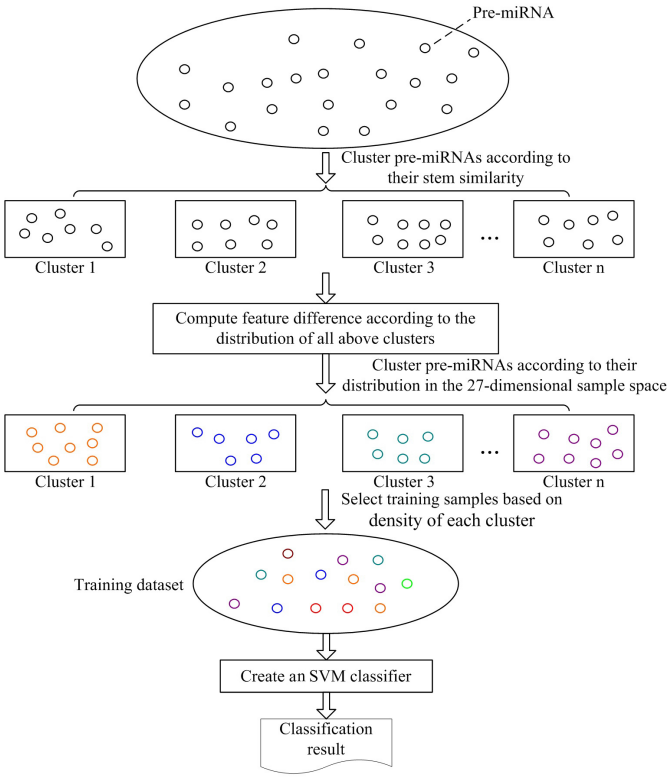


**Figure 1.** Samole selection process based on two-stage clustering.

### *Clustering based on stem similarity*

Since the stems of pre-miRNAs are highly conserved, the nucleotide sequences of stems among the similar pre-miRNAs are usually consistent. We truncate the conserved stems from pre-miRNA hairpins and count how many of the same k-mer sequences there are in the stems of two pre-miRNAs. K-mer restricts special k nucleotides to be adjacent. The pre-miRNAs of human, mouse and rat are aligned with ClustalX 2.0. Most of the nucleotides in the stems are consistent in the 3 species. However, not all the successive nucleotides are consistent. The nucleotides are different at irregular intervals. It is found that there are more same 4-mers in the similar pre-miRNAs from the 3 species. Therefore, 4-mer is selected to measure the stem similarity between two pre-miRNAs.

In the first stage, the pre-miRNA sequences with similar stems are merged into the same cluster. Before clustering based on stem similarity, the appearances of 4-mers should be counted. As shown in Figure 2, given the primary sequences of pre-miRNAs, such as the primary sequence of hsa-mir-192, the secondary structures of the pre-miRNAs are predicted by RNAfold (Hofacker et al., 1994). The central loop and the unpaired part between the 5' arm and 3' arm are then cut off to obtain the conserved stems. Both stems of pre-miRNAs are scanned with a sliding window whose length is 4 nt and the step length is 1 nt. The frequencies of 4-mers in the stems of the 5' arm and 3' arm are counted. For a hairpin with multi-loops, the multi-stems are combined to calculate the stem similarity.

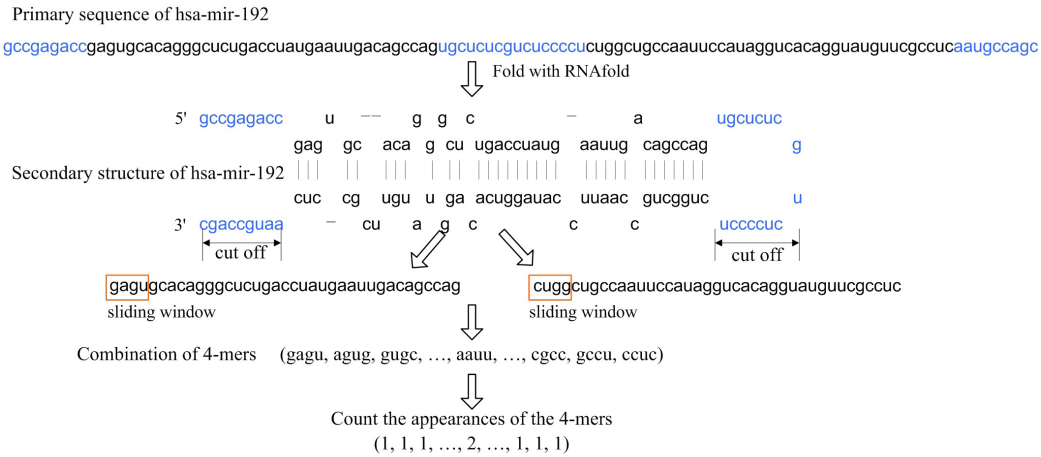


**Figure 2.** Process of counting the appearances of 4-mers.

The cluster algorithm is described as follows, where the threshold $\varepsilon$ is determined by the experiment. When $\varepsilon$ is set to 12, most of the pre-miRNAs with similar stems could be gathered into the same cluster.

Algorithm: Clustering algorithm based on stem similarity.

Input: $N$ real/pseudo pre-miRNAs.

Output: $M$ clusters based on stem similarity.

1. Each real/pseudo pre-miRNA is to be as a single cluster, denoted as $C_1$, $C_2$, …, $C_N$.
2. The similarity between each pair of clusters is calculated.

Suppose two clusters are $C_x$ and $C_y$. $C_x$ is composed of $m$ real/pseudo pre-miRNAs and $C_x = \{S_{x1}, S_{x2}, \ldots, S_{xm}\}$. $C_y$ is composed of $n$ real/pseudo pre-miRNAs and $C_y = \{S_{y1}, S_{y2}, \ldots, S_{yn}\}$. The similarity of $C_x$ and $C_y$ is the number of same 4-mers in all of their stems.

$$Sim(C_x, C_y) = \left| S_{x1} \cap S_{x2} \cap \ldots S_{xm} \cap S_{y1} \cap S_{y2} \cap \ldots S_{yn} \right| \quad \text{(Equation 1)}$$

3. While there still exists $C_x$ and $C_y$ whose $Sim(C_x,C_y) > \varepsilon$
4. do
5. Select two clusters ($C_i$ and $C_j$) with the most similarity and merge them into one cluster $C_k$.
6. Calculate the similarity between the new merged cluster $C_k$ and each of the other clusters again. Thus, the similarity between each pair of clusters is obtained.
7. End While.

## *Feature subset selection*

The positive/negative samples have 48 features, which are extracted from the stem region and the loop region of pre-miRNAs. These features include some unnecessary features, which are useless for classification. Therefore, it is necessary to select discriminative feature subset. Since the stems of pre-miRNAs are highly conservative, the result of first stage clustering reflects the sample distribution based on stem similarity. If a feature could reflect the conservation of stems well, a candidate hairpin sequence with value of the feature similar to one of real pre-miRNAs more likely is a real pre-miRNA. This kind of features should be selected first. The feature difference is defined to measure the average variation of a feature among all the clusters. Here, the feature difference is introduced in this study for the first time.

Suppose $x$ is a feature and the real pre-miRNAs have been gathered into $M$ clusters. $N_i$ is the size of the $i^{th}$ cluster. $v_{ij}$ is the 48-dimensional feature vector of the $j^{th}$ pre-miRNA in the $i^{th}$ cluster. $v_{ij}[k]$ is the $k^{th}$ dimensional feature value of the $j^{th}$ pre-miRNA. The vector set of the $i^{th}$ cluster is $V_i = \{v_{i1}, v_{i2}, \ldots v_{iNi}\}$. The mean value of the $k^{th}$ feature in the $i^{th}$ cluster is $Avg_{ik}$. The root-mean-square value of the $k^{th}$ feature is $DAvg_{ik}$. $MDAvg_k$ represents the average difference value of the $k^{th}$ feature in $M$ clusters.

$$Avg_{ik} = \frac{\sum_{j=1}^{N_i} v_{ij}[k]}{N_i} \quad \text{(Equation 2)}$$

$$DAvg_{ik} = \sqrt{\frac{\sum_{j=1}^{N_i} (v_{ij}[k] - Avg_{ik})^2}{N_i}} \quad \text{(Equation 3)}$$

$$MDAvg_k = \frac{\sum_{i=1}^{M} DAvg_{ik}}{M} \quad \text{(Equation 4)}$$

In addition, in order to better discriminate the positive sample and the negative sample, the information gain of each feature should also be considered. Since all the features of pre-miRNAs are discrete, the feature discrimination is measured by information gain based on Shannon entropy. Suppose a feature of pre-miRNAs is $x$, and the entropy of $x$ is denoted as $H(x)$. When the value of feature $y$ is known, the conditional entropy is $H(x|y)$.

The information gain of feature $x$ and $y$ is $IG(x,y)$ (Quinlan, 1993).

$$IG(x, y) = H(x) - H(x \mid y) \quad \text{(Equation 5)}$$

Classification of real or pseudo pre-miRNAs is a binary class problem. $IG(c, x)$ is the information gain of feature $x$ relative to the classification feature $c$ and $IG(c, x) = H(c) - H(c|x)$. $IG(c, x)$ are used to measure feature discrimination for the training dataset composed of real pre-miRNAs and pseudo pre-miRNAs. The features with greater information gain should be selected first.

However, some features have very small information gain. The features would not improve the classification performance and even have a negative effect on the classifier. Thus, they are useless features and should be avoided.

Since selected features should be consistent with the sample distribution obtained from the first clustering stage, the feature with lower $MDAvg_k$ value is better. Furthermore, the feature with greater information gain should be selected first. The selection weight ($SW$) is used to represent the weight of each feature to be selected. The $SW$ of the $k^{th}$ feature is denoted as $SW_k$. Because the information gain ($IG_k$) is more important than the average feature difference ($MDAvg_k$), $MDAvg_k$ is multiplied by 1/3 to coordinate the proportion between $IG_k$ and $MDAvg_k$. The weight 1/3 is determined according to prior experience (Ng and Mishra, 2007).

$$SW_k = IG[k, c] - MDAvg_k / 3 \quad \text{(Equation 6)}$$

The process of feature selection includes four steps. 1) The feature difference of 48 features is calculated according to the clustering result of the first stage. 2) All the positive samples and negative samples are combined to determine the entropy value and information gain of each feature. 3) The selection weight of each feature is calculated. All values of each feature are normalized to have real values in the interval (0, 1). 4) A threshold is assigned, and only the features with $SW$ greater than the threshold are selected. The threshold is 0.01, which is determined according to prior experience (Sewer et al., 2005; Ng and Mishra, 2007). Table 1 in the section of 'Results and Discussion' shows the selected 27 features and their selection weight.

### *Clustering based on sample distribution*

The first clustering stage gathers the pre-miRNAs with similar stems together, which could be as the initial clusters of second clustering stage. The real/pseudo pre-miRNAs are clustered further according to their position in 27-dimensional sample space. Since the 27 discriminative features are selected, each real/pseudo is represented with a 27-dimensional feature vector. The clustering algorithm of second stage is as follows.

Algorithm: Clustering algorithm based on sample distribution.

Input: $M$ clusters based on stem similarity.

Output: $M$ clusters based on sample distribution.

1. $M$ sample clusters obtained from the first stage are as the initial clusters.
2. The central points of $M$ clusters are calculated, denoted as $m_1, m_2, \ldots, m_M$.
3. While any of $m_1, m_2, \ldots$, and $m_M$ is changed.....
4. Calculate the distance between any one of samples $v$ and $m_1, m_2, \ldots, m_M$, respectively. The distance between $v$ and $m_i$ is $d_{vmi}$, where $v^t$ means the inverse of $v$.

$$d_{vm_i} = 1 - \frac{v^t \cdot m_i}{v^t \cdot v + m_i^t \cdot m_i - v^t \cdot m_i} \quad \text{(Equation 7)}$$

5. $v$ joins the nearest cluster.
6. $m_1, m_2, \ldots$, and $m_N$ are calculated again.
7. End While.

### *Sample selection based on density*

After the second stage clustering, the samples closer to each other in 27-dimensional space are gathered into the same cluster. The training samples are selected according to the density of each cluster. The sample selection process of the $i^{th}$ cluster is as follows.

1. For the central point of the $i^{th}$ cluster, its 27-dimensional feature vector is $m_i$. The number of samples in the $i^{th}$ cluster is $N_i$. $v_k$ is the feature vector corresponding to the $k^{th}$ sample.

$$m_i = \frac{1}{N_i} \sum_{k=1}^{N_i} v_k \quad \text{(Equation 8)}$$

2. The distance between the $k^{th}$ sample and the central point $m_i$ is calculated, denoted as $d_k$. The radius of the $i^{th}$ cluster is $r_i$ and $r_i = max(d_k)$ $(1 \le k \le N_i)$.
3. Suppose the selection rate of sample space is $1/n$. That is, $N_i/n$ samples in the $i^{th}$ cluster would be selected. The number of selected samples is denoted as $P_i = N_i/n$.
4. Suppose $m_i$ is the center of a circle, draw two circles with radius $0r$ and $(1/P_i)r$, respectively. The region between these two circles is $A_0$. The coverage degree of each sample $s$ in $A_0$ is calculated, which is denoted as $C(s)$. $C(s)$ represents the number of samples whose nearest neighbor sample is $s$ in $A_0$. The sample $s$ with the greatest $C(s)$ value is selected as a training sample.
5. We set $(1/P_i)r_i$ as the step length and compute the coverage degree of samples in the region $A_k$ between two circles with the radius $(1/P_i)kr_i$ and $(1/P_i)(k+1)r_i$ $(1 \le k \le P_i - 1)$, respectively. The sample in $A_k$ with greatest coverage degree is selected. The training dataset is composed of all the selected samples.

## Evaluation method

The selected training datasets are used to construct an SVM classifier to evaluate our method. The performance of the classifier is measured with three parameters: the sensitivity ($SE$), the specificity ($SP$), and geometric mean ($G_m$).

$$SE = \frac{TP}{TP + FN}, \quad SP = \frac{TN}{TN + FP}, \quad G_m = \sqrt{SE \times SP} \quad \text{(Equation 9)}$$

where *TP*, *FN*, *TN*, and *FP* denote the number of real/pseudo pre-miRNAs detected/missed, correspondingly. *SE* is the proportion of the positive samples (real pre-miRNAs) correctly classified, and specificity is the proportion of the negative samples (pseudo pre-miRNAs) correctly classified.

## Implementation

Our sample selection method is implemented as *miRNASampleSelect* in Java JDK 1.6. *miRNASampleSelect* can be used in any OS with JVM, including Windows, Linux, Unix, etc. After sample selection, the SVM classifier is created with the libSVM2.9 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). Discretization of feature value could help to compute the entropy and information gain of each feature. We discretize a group of values of each feature with the discretization package supported by Weka 3.7.0.

The *miRNASampleSelect* offers a tool for the high-dimensional pre-miRNAs to select representative training dataset. The selected training dataset contributes to improve the classification performance. The feature vectors corresponding to known real pre-miRNAs and pseudo pre-miRNAs are put into the *miRNASampleSelect* as input. The output of *miRNASampleSelect* is the positive and negative training datasets.

## Complexity analysis

Suppose the size of original sample dataset is $m$ and the average length of real/pseudo pre-miRNAs is $l$. Calculating the appearances of 4-mers takes $O(ml)$. Clustering samples based on the stem similarity needs $O(m)$. Further clustering based on the distribution of samples in high-dimensional space takes $O(m)$. Given the selection rate of sample space is $1/n$, the time cost of selecting samples according to the density of each cluster is $O(m^2/n)$. The total running time for whole algorithm is $O(ml+m+m+m^2/n)$. Since $l<<m$ and $n<<m$, the time complexity of the sample selection is $O(m^2)$.

# RESULTS AND DISCUSSION

## Data collection

A classifier of pre-miRNAs should distinguish real human pre-miRNA hairpins from both pseudo hairpins and other ncRNAs. Therefore, the positive dataset should be composed of known human pre-miRNAs, while the negative dataset should be composed of both pseudo hairpins and human other ncRNAs.

### *Positive dataset*

In order to compare our feature selection method with the *microPred's* method, we use the same positive dataset with *microPred*. Therefore, the dataset includes 695 human pre-miRNAs published in miRBase12.0 (Griffiths-Jones et al., 2008) instead of the current version miRBase15.0. After the redundant sequences have been filtered out, there are 691 non-redundant sequences. Six hundred and sixty of these sequences are folded into hairpin secondary structures and the remaining 31 sequences have multi-branched loops folded with

the RNAfold program. In order to identify multiple types of pre-miRNAs, all of these 691 non-redundant pre-miRNAs are used as positive dataset.

### *Negative dataset*

The 8494 human pseudo hairpin sequences, which are extracted by *triple-SVM* from the protein coding regions, are used as negative dataset. These pseudo pre-miRNAs have been previously used in *triple-SVM*, *MiPred*, *miPred*, and *microPred*. The criteria of selecting the pseudo-miRNAs are: minimum of 18 base pairings on the stem of hairpin structure, maximum of -15 kcal/mol free energy of secondary structure, and no multiple loops, which ensures that the extracted pseudo pre-miRNAs are similar to real pre-miRNAs. In addition, the negative dataset also includes 754 other ncRNAs collected by *microPred*, where some ncRNAs have multiple loops, total 9248 sequences.

### *Positive and negative training dataset*

With the sample selection method, 311 pre-miRNAs are selected from 691 known real pre-miRNAs as the positive training set and 411 pseuso pre-miRNAs are selected from 9248 pseudo pre-miRNAs as the negative training set. The selected training dataset is referred to as "722 training dataset".

### *Positive and negative testing dataset*

Two groups of positive and negative testing dataset are created. The first group is composed of 350 real pre-miRNAs and 350 pseudo pre-miRNAs. The 350 real pre-miRNAs are randomly selected from the remaining positive dataset excluding the 311 real pre-miRNAs. The 350 pseudo pre-miRNAs are selected from 8494 pseudo pre-miRNAs excluding the 411 pseudo pre-miRNAs. The first group of testing samples is referred to as "700 random testing dataset". In the second group, the positive testing dataset is composed of 691 known real pre-miRNAs and the negative dataset consists of 754 ncRNAs. The second group of testing samples is referred to as "1445 real and ncRNA testing dataset". It is well known that some ncRNAs are often wrongly classified as real pre-miRNAs for many classifiers. Therefore, all the ncRNAs are added to the second negative testing dataset.

## Feature subset selection result

Each real/pseudo pre-miRNA originally has 48 features, which include some useless features. Therefore, the most discriminative feature subset is selected, which contributes to the better description of the distribution of real/pseudo pre-miRNAs in sample space. The selected 27 features, the corresponding information gain, the MDAvg, and the selection weight are shown in Table 1, which are sorted according to their selection weight.

We found that 16 dinucleotide frequencies (*AA%,AC%,AG%,...,UU%*) were nearly useless and they were not selected by the feature selection method. There is a strong consensus result in *miPred* that indirectly confirms our selected feature subset. It is well studied that the stem-loop structures of pre-miRNAs is thermodynamically stable. Most of the selected features are related to the thermodynamic stability of the secondary structures. It further confirms the effectiveness of the selected features.

**Table 1.** Selected features ranked according to their selection weight.

| Rank | AttrName | IG (c, attr) | MDAvg | SW | Rank | AttrName | IG (c, attr) | MDAvg | SW |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Freq | 1 | 0.295 | 0.901 | 15 | \|A-U\|/L | 0.317 | 0.251 | 0.233 |
| 2 | Diversity | 0.779 | 0.297 | 0.681 | 16 | %(A-U)/stems | 0.232 | 0.078 | 0.206 |
| 3 | MFEI1 | 0.647 | 0.075 | 0.621 | 17 | Tm | 0.303 | 0.305 | 0.202 |
| 4 | ZG | 0.633 | 0.181 | 0.573 | 18 | EAFE | 0.264 | 0.251 | 0.181 |
| 5 | dP | 0.504 | 0.264 | 0.415 | 19 | ZF | 0.249 | 0.315 | 0.143 |
| 6 | ZP | 0.471 | 0.247 | 0.388 | 20 | dF | 0.161 | 0.103 | 0.126 |
| 7 | ZQ | 0.401 | 0.117 | 0.362 | 21 | dH/L | 0.154 | 0.102 | 0.119 |
| 8 | dQ | 0.376 | 0.132 | 0.332 | 22 | dH | 0.158 | 0.207 | 0.089 |
| 9 | Avg_Bp_Stem | 0.358 | 0.114 | 0.320 | 23 | MFEI4 | 0.131 | 0.137 | 0.085 |
| 10 | ZD | 0.358 | 0.130 | 0.314 | 24 | Diff | 0.136 | 0.226 | 0.061 |
| 11 | MFEI3 | 0.316 | 0.039 | 0.303 | 25 | dS/L | 0.097 | 0.131 | 0.054 |
| 12 | MFEI2 | 0.301 | 0 | 0.301 | 26 | dS | 0.116 | 0.205 | 0.047 |
| 13 | dG | 0.317 | 0.085 | 0.288 | 27 | %G+C | 0.100 | 0.263 | 0.012 |
| 14 | dD | 0.338 | 0.166 | 0.283 | | | | | |

When the real/pseudo pre-miRNAs are represented by different feature subsets, the sample selection algorithm would select different training datasets. The classification performances of the classifiers learning from these different training datasets are compared, as shown in Table 2. Three feature subsets include 21 features selected by *microPred* with *J-M* Distance, 20 features with smaller feature difference, and 27 features selected based on selection weight. For each feature subset, 311 real pre-miRNAs and 411 pseudo pre-miRNAs are selected by the sample selection method to train an SVM classifier. Three SVM classifiers are tested on the ***700 random testing dataset***. Obviously, the selected feature subset based on selection weight achieves better classification performance compared to the other two feature subsets. This further confirms the importance of the feature selection in the sample selection.

**Table 2.** Classification result with different feature subsets.

| Feature selection method | Number of selected features | Classification results (%) | | |
|---|---|---|---|---|
| | | *SE* | *SP* | $G_m$ |
| *J-M* (*microPred*) | 21 | 91.69 | 93.25 | 92.47 |
| Feature with smaller feature difference | 20 | 83.95 | 97.11 | 90.29 |
| Feature selection based on selection weight | 27 | 99.74 | 99.74 | 99.74 |

*SE* = sensitivity; *SP* = specificity; $G_m$ = geometric mean.

## Training sample selection result

The selected training samples with the *miRNASampleSelect* are used to create the SVM classifier, referred to as *miRNAPred*. The performances of classifiers trained with different sample selection methods, including *SMOTE* in *microPred* and the *miRNASampleSelect*, are compared.

As shown in Table 3, 5-fold cross-validation is performed on the training data to compare the performance of two classifiers. We performed 10 repeated evaluations for each testing dataset and averaged the results. The result of 5-fold cross-validation with *SMOTE* is obtained from the publication on *microPred*. Other testing results of *microPred* are obtained through accessing the web server of *microPred* (http://web.comlab.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm). The experimental results indicate that the classifier *miRNAPred* outperforms *microPred* significantly. First, *SE* increased by 9.5% on average. The improvement of *SE* could be of benefit for detecting more new pre-miRNAs. Second, *SP* increased by 14.82%

on average. Specificity is helpful to greatly decrease false predictions because of the large size of genome sequences. Therefore, the improvement of *SP* is a very significant increase. *miRNAPred* is nearly 12% greater in total accuracy. Thus, *miRNAPred* achieves higher and, especially, much more reliable classification results than *microPred* in terms of both sensitivity and specificity.

**Table 3.** Classification results with different sample selection methods.

| Sample selection method | Dataset | Classification results (%) | | |
|---|---|---|---|---|
| | | *SE* | *SP* | $G_m$ |
| *SMOTE* (5-fold cross-validation) | Training dataset generated by microPred | 90.02 | 97.28 | 93.58 |
| *miRNASampleSelect* (5-fold cross-validation) | 722 training dataset | 99.40 | 99.34 | 99.37 |
| *SMOTE* | 700 random testing dataset | 90.00 | 77.43 | 83.48 |
| *miRNASampleSelect* | | 99.71 | 99.14 | 99.42 |
| *SMOTE* | 1445 real and ncRNA testing dataset | 90.16 | 77.59 | 83.64 |
| *miRNASampleSelect* | | 99.57 | 98.28 | 98.93 |

*SE* = sensitivity; *SP* = specificity; $G_m$ = geometric mean.

Almost all the pre-miRNAs with multiple loops in the testing dataset could be classified correctly, which indicates that, unlike previously reported methods, our method could be sensitive enough to identify pre-miRNAs with multi-loops. There are 4 pre-miRNAs that are easily misjudged in the positive testing dataset composed of 691 known pre-miRNAs. There are multiple big loops in the precursor of hsa-mir-375. All the precursors of hsa-mir-1308, hsa-mir-1469, and hsa-mir-1825 only include 15 bp. Thus, their secondary structures are not stable enough. The description above might be the reason that our classifier could not classify these 4 pre-miRNAs correctly.

## CONCLUSION

We investigated a novel sample selection method (*miRNASampleSelect*) according to the characteristics of pre-miRNAs. *miRNASampleSelect* effectively solved the class imbalance problem in classification of real pre-miRNAs and pseudo pre-miRNAs. The classifier *miRNAPred* trained with the selected samples achieved higher sensitivity and specificity.

The 27 discriminative features are selected to describe the distribution of real/pseudo pre-miRNAs. These features could also be used in the other kinds of classification models, such as naive Bayes and random forest, which contribute to improving their classification performance. The 311 real pre-miRNAs and 411 pseudo pre-miRNAs are selected as training samples. These training samples could be used to train other SVM classifiers, such as *triplet-SVM*, *MiPred*, *miPred*, and *microPred*, to increase their classification accuracy.

## ACKNOWLEDGMENTS

Research supported in part by the Returned Scholar Foundation of Educational Department of Heilongjiang Province in China, under grant #1154hz26.

## REFERENCES

Akbani R, Kwek S and Japkowicz N (2004). Applying Support Vector Machines to Imbalanced Datasets. In: Machine Learning: ECML 2004 (Boulicaut JF, Esposito F, Giannotti F and Pedreschi D, eds.). Springer Berlin/Heidelberg, 39-50.

Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116: 281-297.
Batuwita R and Palade V (2009). microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* 25: 989-995.
Berezikov E, Guryev V, van de Belt J, Wienholds E, et al. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 120: 21-24.
Bushati N and Cohen SM (2007). microRNA functions. *Annu. Rev. Cell Dev. Biol.* 23: 175-205.
Chang DT, Wang CC and Chen JW (2008). Using a kernel density estimation based classifier to predict species-specific microRNA precursors. *BMC Bioinformatics* 9 (Suppl 12): S2.
Chatterjee S and Grosshans H (2009). Active turnover modulates mature microRNA activity in *Caenorhabditis elegans*. *Nature* 461: 546-549.
Chawla NV, Browyer KW, Hall LO and Kegelmeyer WP (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16: 321-357.
Fera D, Kim N, Shiffeldrim N, Zorn J, et al. (2004). RAG: RNA-As-Graphs web resource. *BMC Bioinformatics* 5: 88.
Freyhult E, Gardner PP and Moulton V (2005). A comparison of RNA folding measures. *BMC Bioinformatics* 6: 241.
Gan HH, Fera D, Zorn J, Shiffeldrim N, et al. (2004). RAG: RNA-As-Graphs database - concepts, analysis, and features. *Bioinformatics* 20: 1285-1291.
Griffiths-Jones S, Saini HK, van Dongen S and Enright AJ (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36: D154-D158.
Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, et al. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125: 167-188.
Imam T, Ting K and Kamruzzaman J (2006). z-SVM: An SVM for Improved Classification of Imbalanced Data. In: Proceedings of the 19th Australian Joint Conference on Artificial Intelligence (AI 2006): Advances in Artificial Intelligence (Sattar A and Kang BH, eds.). Springer, Berlin/Heidelberg, 264-273.
Jiang P, Wu H, Wang W, Ma W, et al. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35: W339-W344.
Molinara M, Ricamato MT and Tortorella F (2007). Facing Imbalanced Classes through Aggregation of Classifiers. In: Proceeding of the 14th International Conference on Image Analysis and Processing, 10-14 September, IEEE Computer Society, Modena, 43-48.
Moulton V, Zuker M, Steel M, Pointon R, et al. (2000). Metrics on RNA secondary structures. *J. Comput. Biol.* 7: 277-292.
Nam JW, Shin KR, Han J, Lee Y, et al. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* 33: 3570-3581.
Ng KL and Mishra SK (2007). *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics* 23: 1321-1330.
Quinlan JR (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Mateo.
Schultes EA, Hraber PT and LaBean TH (1999). Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.* 49: 76-83.
Seffens W and Digby D (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* 27: 1578-1584.
Sewer A, Paul N, Landgraf P, Aravin A, et al. (2005). Identification of clustered microRNAs using an *ab initio* prediction method. *BMC Bioinformatics* 6: 267.
Veropoulos K, Campbell C and Cristianini N (1999). Controlling the Sensitivity of Support Vector Machines. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99). Morgan Kaufmann, Stockholm, 55-60.
Wang J, Zou Q and Guo MZ (2010). Mining SNPs from EST sequences using filters and ensemble classifiers. *Genet. Mol. Res.* 9: 820-834.
Weiss GM (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations Newsl.* 6: 7-10.
Xue C, Li F, He T, Liu GP, et al. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 6: 310.
Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, et al. (2006). Combining multi-species genomic data for microRNA identification using a naive Bayes classifier. *Bioinformatics* 22: 1325-1334.
Yousef M, Jung S, Showe LC and Showe MK (2008). Learning from positive examples when the negative class is undetermined - microRNA gene identification. *Algorithms Mol. Biol.* 3: 2.
Zhang BH, Pan XP, Cox SB, Cobb GP, et al. (2006). Evidence that miRNAs are different from other RNAs. *Cell Mol. Life Sci.* 63: 246-254.