

# Analysis of miRNA expression profile based on SVM algorithm

Dai Ting-ting<sup>1, a</sup>, Shan Chang-ji<sup>2, b</sup>, Dong Yan -shou<sup>3, c</sup>, Bian Yi-duo<sup>4, d</sup>

<sup>1</sup>School of mathematics and statistics, Zhaotong University, Yunnan, 657000, China,

<sup>2</sup>School of physics and information engineering, Zhaotong University, Yunnan, 657000, China,

<sup>3</sup>school of foreign languages, Zhaotong University, Yunnan, 657000, China

<sup>a</sup>876310867@qq.com, <sup>b</sup>shanchangji@126.com

**Abstract.** Based on mirna expression spectrum data set, a new data mining algorithm - tSVM - KNN (t statistic with support vector machine - k nearest neighbor) is proposed. the idea of the algorithm is: firstly, the feature selection of the data set is carried out by the unified measurement method; Secondly, SVM - KNN algorithm, which combines support vector machine (SVM) and k - nearest neighbor (k - nearest neighbor) is used as classifier. Simulation results show that SVM - KNN algorithm has better classification ability than SVM and KNN alone. Tsvm - KNN algorithm only needs 5 mirnas to obtain 96.08 % classification accuracy in terms of the number of mirna " tags" and recognition accuracy. compared with similar algorithms, tsvm - KNN algorithm has obvious advantages.

## 1.Introduction

Mirnas are a class of non-protein encoded single-stranded RNA molecules with regulatory effects, It regulates the expression of one-third of messenger RNA in humans<sup>[1]</sup>, whereas one mirna is abnormally expressed. This leads to differential expression of the corresponding proteins. Therefore, miRNAs are important for tumor formation. In recent years, researchers have discovered that miRNA expression profiles can distinguish between tumor and normal tissues. In addition, studies have shown that some miRNAs have been directly involved in human cancer<sup>[2]</sup>. However, it is still unclear which specific miRNAs can accurately distinguish between normal tissues and tumor tissues. If we can identify some miRNAs "tags" with strong classification ability through specific methods, then according to the principle of bioinformatics, Combining relevant software to predict the target genes corresponding to these "tags", then it is possible to find new targets for the diagnosis and treatment of diseases such as tumors. Therefore, the analysis of miRNAs expression profile data in clinical studies has a greater diagnosis value.

In the literature<sup>[3]</sup>, Zheng et al. used a discrete function learning algorithm to find a subset of miRNAs with strong classification ability. In 2010, Dang et al<sup>[4]</sup>. proposed two steps using existing data sets. Selecting the feature method yielded a classification accuracy of about 95%. In the same year, Kyung et al<sup>[5]</sup>. used a cosine coefficient as a feature selection method to select 25 features and combined the K-nearest neighbor method as a classifier to obtain a classification accuracy of 95%.

In fact, there are two main stages in the study of tumor classification. The first stage is to find effective feature selection methods; the second stage is to search for efficient classifiers. However, there are many methods for feature selection, such as the rank sum test and the Fisher criterion method,



which have achieved good results. However, the classifier algorithm is relatively single, such as SVM or kNN, so this article attempts to improve the classification accuracy by starting with the classifier algorithm. Based on this, this paper proposes a data mining algorithm-tSVM-kNN. The idea of this algorithm is: First, we use statistical methods to perform feature primaries on the data set. Secondly, we use the SVM-kNN algorithm that combines support vector machines and k-nearest neighbors discriminant method as the classifier, and finally output the classification results.

## 2. Problem description

First, the miRNA expression profile data is expressed in matrix M as:

$$M = \begin{bmatrix} \overbrace{x_{1,1} \ x_{1,2} \ \cdots \ x_{1,n}}^{n \uparrow \text{miRNAs}} & \overbrace{y_1}^{\text{class}} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} & y_m \end{bmatrix}$$

## 3. Methods and Theory

### 3.1. Feature selection<sup>[6]</sup>

The t-test is a parametric test method based on the premise that the sample obeys the assumption of normal distribution, and the t-statistic and its deformation are now commonly used as a kind of tumor identification measure. The t statistic is:

$$t = \frac{\mu_+(i) - \mu_-(i)}{s_w^2 \sqrt{\frac{1}{n_+} + \frac{1}{n_-}}} \quad (1)$$

$$s_w^2 = \frac{(n_+ - 1)\sigma_+^2(i) + (n_- - 1)\sigma_-^2(i)}{n_+ + n_- - 2} \quad \mu_+(i), \mu_-(i), \sigma_+(i), \sigma_-(i) \text{ Represents the}$$

Among them, average and standard deviation of the expression levels of the ith miRNA in positive and negative samples, respectively And the number of positive and negative samples respectively. The greater the absolute value of the t statistic, the more significant the difference in positive and negative samples of the miRNA expression level. Of course, in addition to the t-test method, feature selection methods include rank sum tests and Fisher criterion methods.

### 3.2. Introduction to SVM-kNN Improved Algorithm

Nearest neighbor discriminant is one of the most important methods in pattern recognition nonparametric method. the idea of this algorithm is simple and intuitive [7]. the distance between the sample to be classified and all samples is calculated. considering the category of the sample nearest to the sample to be classified, the sample to be classified also belongs to this category. The characteristic of 1nn is that all the sample points in each class are taken as representative points. k - nearest neighbor discriminant analysis ( KNN ) is a generalization of 1nn. its idea is to investigate the categories of k samples nearest to the sample to be classified. if most of the samples belong to a certain category, the sample to be classified also belongs to that category.

In fact, the SVM-kNN combinatorial algorithm can be described as: When the distance between the sample and the SVM optimal hyperplane is greater than the given threshold, that is, the sample is far from the interface, use the SVM to classify; instead, use the support vector as the representative point.

The kNN algorithm classifies the identified samples. It is worth noting that we calculate the distance between the sample to be identified and each support vector machine in the mapped feature space, rather than in the original sample space. Therefore, the distance formula is used. Instead of the Euclidean distance formula, it is the following formula:

$$d(x, x_i) = \|\phi(x) - \phi(x_i)\|^2 = k(x, x) - 2k(x, x_i) + k(x_i, x_i) \quad (2)$$

Among them,  $x$  want to identify the sample;  $x_i$  is the support vector.

The following is to build the SVM-kNN classifier. First select the appropriate SVM classifier to perform machine learning for a given set of data samples. Finding Support Vector Sets  $T_{SV}$  And the corresponding *Lagrange* Coefficients and constants  $b$ .

SVM-kNN algorithm:

Enter: Test set  $T$ , training set  $Z$ , support vector set  $T_{SV}$ , corresponding *Lagrange* coefficients  $\alpha^{SV}$  and constants  $b$ , coefficient  $k$  of kNN algorithm, given classification threshold  $\varepsilon$ , kernel function and its parameters;

Output: Test set  $T$  classification results;

Start

Step1: If  $T \neq \emptyset$  then  $x \notin T$ , if  $T = \emptyset$  The algorithm stops and outputs the category vector.

Step2: Test sample point  $x$  Into the formula  $g(x) = \sum_{i=1}^n a_i y_i K(x_i, x) + b$  Calculate the distance from

the sample point to the optimal classification surface

Step3: if  $|g(x)| \geq \varepsilon$ , Then call the *SVM* algorithm, Calculation  $f(x) = \text{sgn}(g(x))$  As output;

Step4: if  $|g(x)| < \varepsilon$  Then call the *kNN* algorithm, The specific operation is as follows:

- 1) Calculate the distance to all support vectors, Find  $d_{sv} = \|\phi(x) - \phi(x_{sv})\|^2$
- 2) Sort these distances from small to large and choose the smallest  $k$  numbers.
- 3) Count the number of categories of support vectors corresponding to this  $k$  distance.
- 4) The category is the same as the number of categories and the output.

Step5:  $T \leftrightarrow T - \{x\}$ , return Step1.

The kernel function used in the above algorithm can be selected according to the actual problem. The classification threshold can also be dynamically adjusted according to the experimental results. Its initial value is generally set to 1. If and only then, the above algorithm completely degenerates into the SVM algorithm.

### 3.3. Cross -validation

Cross-validation is a statistical analysis method used to verify the performance of classifiers. The basic idea is to group the original data sets in a certain sense, one part as a training set and the other part as a verification set. First, train the classifier with the training set, and then use the verification set to test the trained model as a performance indicator for evaluating the classifier. This article uses a 10-fold cross-validation method to evaluate the performance of the *tSVM-kNN* algorithm.

## 4. Empirical analysis

### 4.1. Data sources and processing

The experimental data used in this paper was from the miRNA expression profile data (miGCM\_218collection) published by Lu et al<sup>[8]</sup>. and the data can be downloaded from the website <http://www.broadinstitute.org>. The dataset contains a total of 186 samples and contains multiple cancer types, and each sample contains 217 miRNAs expression data. In addition, there are 46 positive

samples and 140 negative samples in the dataset. In order to eliminate the dimensional relationship between variables and Makes the data comparable, prior to data analysis, Z-score normalization of sample data. Then the positive and negative samples were randomly assigned to the training set and test set in a ratio of approximately 5:2. According to the miRNAs feature selection method, the complete code of the miRNA expression profile analysis process was implemented. The tool used was Matlab2012a.

#### 4.2. Feature primary

Since the purpose of the experiment is to select a small number of features and achieve a good classification effect, the SVM-KNN classifier classification effect of 2-8 features is calculated using the statistical method (see Table 1). In addition, the classification threshold in the corresponding SVM-KNN classifier is also given in Table 1.

Table 1 Comparison of t-statistics features

Method	Features' numbers						
	2	3	4	5	6	7	8
The Method of t statistic	92.16% $\varepsilon = 0.3$	92.16% $\varepsilon = 0.4$	94.12% $\varepsilon = 0.2$	96.08% $\varepsilon = 0.3$	94.12% $\varepsilon = 0.6$	94.12% $\varepsilon = 0.5$	92.16% $\varepsilon = 0.4$

From Table 1, we can see that for the statistical method, as long as the top five features are selected, a good classification effect can be achieved. The features selected by the statistics method are listed, as shown in Table 2.

Table 2 The top 5 features selected by the t-statistic

Hsa-miR-195:	UAGCAGCACAGAAAUUUGGC:bead_165-A
Hsa-miR-101:	UACAGUACUGUGAUGACUAAG:bead_163-B
Hsa-miR-126:	UCGUACCGUGAGUAAUAAUGC:bead_166-B
Hsa-miR-26a	UCGUACCGUGAGUAAUAAUGC:bead_112-B
rno-miR151*:	UCGAGGAGCUCACAGUCUAGU:bead_160-C

#### 4.3. Determination of parameters $c, g$ and $k$ in SVM-kNN model

When using the SVM (RBF kernel) classifier for classification, this paper uses the 10-fold cross-validation to select the optimal penalty parameters  $C=0.1088$  and  $g=0.0625$ .

When calling the kNN classifier, the value of  $k$  should be between 1 and the number of support vectors, and is an odd number (to avoid appearance). This article uses statistics to perform feature primaries, and then puts the selected feature subset into the SVM machine. Learning, the number of support vectors obtained is 71, therefore, the range of values is an odd number between 1 and 71. In this paper, the determination of the optimal parameter  $k$  depends on the  $k$  value corresponding to the maximum classification accuracy of the SVM-kNN classifier.

#### 4.4. Result analysis

Table 3 describes a comparison of the classification accuracy of the SVM-kNN combined model with that of a single SVM classifier and a kNN classifier. It can be seen that the classification effect of the SVM-kNN classifier is better than running the SVM classifier and kNN class separately. The classification effect of the device is better. Among them, tSVM-kNN has a classification accuracy of 96.08%.

Table 3 Classification Accuracy under Different Classification Models

The Method of Feature selection	Classifier	Classification accuracy
The Method of $t$ statistic	kNN( $k=69$ )	88.24%
The Method of $t$ statistic	SVM(RBF kernel)	88.24%
The Method of $t$ statistic	SVM-kNN( $k=5$ )	96.08%

#### 4.5. Comparison

In recent years, many researchers have devoted themselves to the analysis of miRNA expression profile data, and hope to use "potential miRNAs" (ie, miRNAs "tags") to find potential therapeutic targets for tumors. Therefore, it is possible to find the ability to classify tumors. It is particularly important to better distinguish between positive and negative samples of miRNAs. Table 4 shows the comparison of experimental results using different feature selection methods and different classifiers for the same data set. The experimental results show that the method of this paper is "labeled" in miRNAs. The number and recognition accuracy have obvious advantages. At the same time, the tSVM-kNN algorithm has stronger competitive advantages with other algorithms. Its biggest advantage is reflected in the classifier SVM-kNN.

**Table 4 Comparison of experimental results of different classification methods for the same data set**

The Method of	Classifier	Classification	The number of	Reference
Feature selection		accuracy	Informative miRNAs	[9]
The Method of $t$ statistic	SVM-kNN	96.08%	5	[9]
Cosine coefficients	kNNS	95.00%	25	[9]
Pearson correlation	Multi-Layer Perceptron	92.50%	25	[9]
Spearman correlation	SVM(linear)	91.7%	25	[9]
Mutual information	kNNE	93.00%	25	[9]
Signal-to-noise ratio	SVM(linear)	90.60%	25	[9]
Information gain	kNNP	92.9%	25	[9]
Two-step feature selection method	SVM(linear)	95.00%	20	[8]

## 5. Conclusions and Prospects

Based on the dataset of miRNAs expression profiles, this paper proposes a data mining method named tSVM-kNN which uses statistic method as feature selection method and SVM-kNN algorithm as classifier. Experimental results show that the SVM-kNN algorithm classifier has better classification ability than SVM and kNN alone. In terms of the number of miRNA "tags" and knowledge accuracy, the tSVM-kNN algorithm requires only 5 miRNAs to obtain 96.08%. Classification accuracy. Compared with similar algorithms, it has obvious advantages.

## References

- [1] LEA M A. recently identified and potential targets for colon cancer treatment[J]. Future Oncology,2010,6(6):993.
- [2] ZHANG B, PAN X, COBB G P et al. MicroRNAs as oncogenes and tumor suppressors[J]. Developmental Biology,2007,302(1):1.
- [3] DALMAY T. MicorRNAs and cancer [J]. Journal of Internal Medicine,2008,263(4):366.
- [4] WU W, SUM M, ZOU G M, et al. MicorRNAs and cancer: Current status and prospective[J]. International Journal of Cancer,2007,120(5):953.
- [5] DRAKNKI A, ILIOPOULOS D. MicroRNA genenetworks in oncogenesis[J]. Current Genomics,2009,10(1):35.
- [6] LU J, GETZ G, MISKA E A, et al. MicroRNA expression profiles classify human cancers [J]. Nature,2005,435:834.
- [7] ZEHNG Y, KWOH C K. Cencer classification with microRNA expression patterns found by an information theory approach [J]. Journal of Computers,2006,1(5):30.
- [8] TRNA D H, HO, T B, PHAM T H, et al. MicroRNA expression profiles for classification and analysis of tumor samples[J]. IEICE Transactons on Information and Systems,2011,94(3):416.