Exploring Features and Classifiers to Classify MicroRNA Expression Profiles of Human Cancer

Kyung-Joong Kim[1] and Sung-Bae Cho[2]

[1] Dept. of Computer Engineering, Sejong University, Seoul, South Korea
[2] Dept. of Computer Science, Yonsei University, Seoul, South Korea
kimkj@sejong.ac.kr, sbcho@cs.yonsei.ac.kr

Abstract. Recently, some non-coding small RNAs, known as microRNAs (miRNA), have drawn a lot of attention to identify their role in gene regulation and various biological processes. The miRNA profiles are surprisingly informative, reflecting the malignancy state of the tissues. In this paper, we attempt to explore extensive features and classifiers through a comparative study of the most promising feature selection methods and machine learning classifiers. Here we use the expression profile of 217 miRNAs from 186 samples, including multiple human cancers. Pearson's and Spearman's correlation coefficients, Euclidean distance, cosine coefficient, information gain, mutual information and signal to noise ratio have been used for feature selection. Backpropagation neural network, support vector machine, and *k*-nearest neighbor have been used for classification. Experimental results indicate that *k*-nearest neighbor with cosine coefficient produces the best result, 95.0% of recognition rate on the test data.

Keywords: microRNA, Human Cancer, Classification, Feature Selection, Machine Learning.

1 Introduction

High-throughput messenger RNA (mRNA) expression profiling with microarray has produced huge amount of information useful for cancer diagnosis and treatment [1]. It has also promoted the development of techniques to analyze the large amount of information using statistical and machine learning approaches [2]. Computational methods selects relevant subsets of thousands genes and classify samples into normal or tumor tissues. Clustering technology reveals the relevant modules of co-expressed genes that show similar behavioral patterns in gene regulation process [3]. There are several microarray databases accessible by public [4][5].

Recently, small-non-coding RNAs, named microRNAs (miRNA) have drawn a lot of attention to identify their functional roles in biological processes [6][7]. Especially, researchers have investigated that the abnormal expression of miRNAs may indicate human diseases, such as cancers. Lu *et al*. collected 217 miRNAs expression profiles from 334 human and mouse samples using a bead based flow cytometric method [8]. They reported a down-regulation of miRNAs in cancer tissues compared with normal ones. In addition to the observation, they applied simple classification algorithms to

K.W. Wong et al. (Eds.): ICONIP 2010, Part II, LNCS 6444, pp. 234–241, 2010.

the samples which are not easily discriminated with mRNA expression profiles. Some researchers have been attempting to propose the optimal classification technique to work out this problem, especially dealing with predictive discrimination of multiple cancers [9][10].

Although there have been several comprehensive works to compare the possible methods with different feature selection and classification techniques for mRNA expression profiles [11], there have been still no work on the miRNA data. Like mRNA classification problems, there are a lot of possible choices on the combination of feature selection methods and classification algorithms resulting in different recognition accuracy. A through effort helps to find the best possible methods to classify human cancer using miRNA expression profiles. Also, it reveals the superiority of specific feature selection method and classification algorithm over alternatives for the problem.

In this paper, we attempt to explore the features and classifiers that efficiently detect the malignancy status (normal or cancer) of the tissues. We have adopted seven feature selection methods widely used in pattern recognition fields: Pearson's and Spearman's correlations, Euclidean distance, cosine coefficient, information gain and mutual information and signal-to-noise ratio. We have also utilized four *k*-nearest neighbor methods with different similarity measures (Euclidean, Pearson and Spearman correlation, and cosine coefficient), multilayer perceptrons, and support vector machines with linear kernel.

2 MicroRNA

Recently, hundreds of small, non-coding miRNAs have been discovered [7] which are averaging approximately 22 nucleotides in length (Table 1). They are involved with cell proliferation and death, gene regulatory networks, RNA metabolism, auxin signaling and neuronal synapse formation [6][7]. Especially, the expression of miRNAs indicates human diseases such as cancers [8]. Lu *et al*. used *k*-nearest neighbor and probabilistic neural network to classify human cancer using miRNA expression profiles. In their work, they used human miRNA expression data for multiple cancers as training samples to predict the mouse lung cancer's malignancy. They reported 100% accuracy for 12 mouse lung cancer tissues.

Table 1. Examples of miRNA expression profiles [8]

Description	Sample 1	Sample 2
hsa-miR-124a:UUAAGGCACGCGGUGAAUGCCA:bead_101-A	7.4204	6.931
hsa-miR-125b:UCCCUGAGACCCUAACUUGUGA:bead_102-A	10.8391	11.7231
hsa-miR-7:UGGAAGACUAGUGAUUUUGUU:bead_103-A	6.64631	6.78163
hsa-let-7g:UGAGGUAGUAGUUUGUACAGU:bead_104-A	9.86267	10.4861
hsa-miR-16:UAGCAGCACGUAAAUAUUGGCG:bead_105-A	10.6879	11.5479
hsa-miR-99a:AACCCGUAGAUCCGAUCUUGUG:bead_107-A	8.39361	8.88749
hsa-miR-92:UAUUGCACUUGUCCCGGCCUGU:bead_108-A	8.63981	9.06636

Xu *et al*. applied a neural based classifier, Default ARTMAP, to classify broad types of cancers based on their miRNA expression profiles [9]. In their work, particle swarm optimization (PSO) was used for selecting important miRNAs that contribute to the discrimination of different cancer types. Zheng *et al*. reported that discrete function learning algorithm (DFL) obtains better prediction performance than C4.5 (decision tree) and RIP algorithms.

3 MicroRNA Expression Classification

Generally, the miRNA expression profile has high dimensionality with small number of samples because of the limitation of sample availability, cost, or other reasons. After acquiring the miRNA expression profile, prediction systems goes through two stages: feature selection and pattern classification stages. The high dimensionality is one of the major challenges to analyze the miRNA expression profiles decreasing classification accuracy. There are a lot of feature selection methods proposed based on statistical similarity, information theory and signal-to-noise ratio [12]. Feature selection methods select relevant miRNAs which contribute to the discrimination of the malignancy type. At prediction stage, classification algorithms learn models with the selected miRNAs to predict the category of each sample. Finally, their goodness of the classifier is evaluated on unseen samples, called test data.

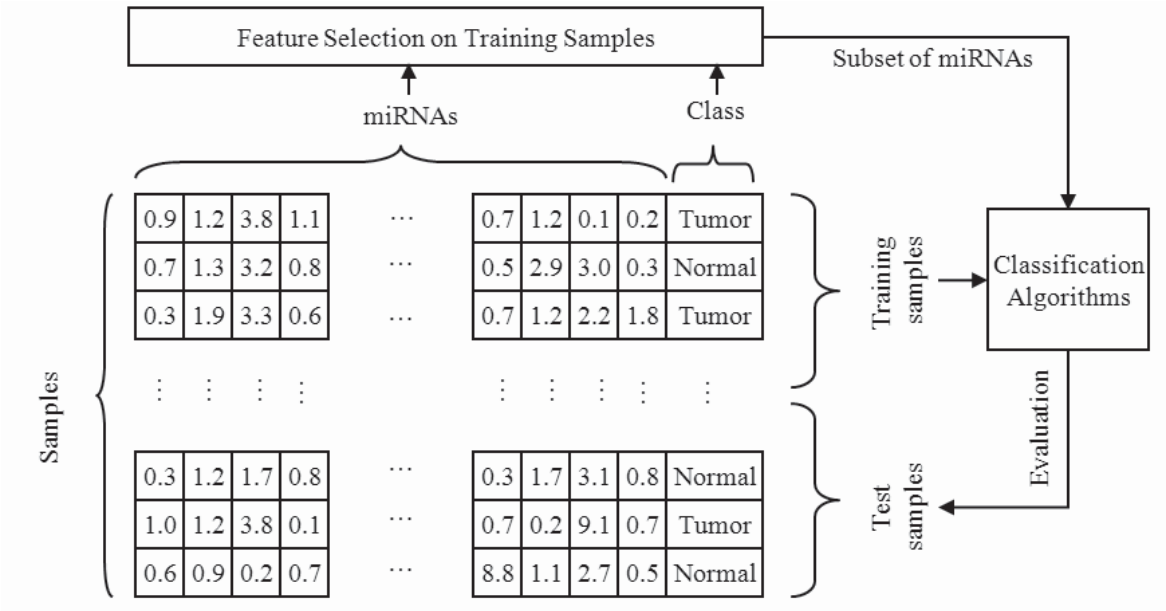


Fig. 1. Overview of miRNA classification system

3.1 Features

3.1.1 Similarity-Based Methods

In these methods, the value of each miRNA is evaluated based on the similarity to ideal vectors. In case of positive ideal vectors, the value is 1 if the training sample is cancer and vice versa. On the other hand, in the case of negative ideal vectors, the

value is 1 if the training sample is normal. If there is a miRNA that shows the same behavior with the ideal vectors, this means that we can classify the training samples correctly with only the single miRNA. Because it is not common to classify samples correctly using only single miRNA, this vector is called as "ideal" one.

We can sort the miRNAs in accordance with the similarity between the miRNA's values for training samples and ideal vectors. Because we have the two ideal vectors, there are two different rankings based on positive and negative ideal vectors. Finally, half of the miRNAs are chosen from the rankings by the positive ideal vector, and others are from the one by the negative ideal vector. For example, if we decide to select 20 miRNAs, 10 miRNAs are very close to the positive ideal vectors and 10 miRNAs are very close to the negative ones. There are four different similarity measures used: inverse of Euclidean distance measure, Pearson correlation, cosine coefficient and Spearman correlation.

3.1.2 Information Gain

In the following formula, k is the total number of classes, n_l is the number of values in the left partition, n_r is the number of values in the right partition, l_i is the number of values that belong to class i in the left partition, and r_i is the number of values that belong to class i in the right partition. The information gain of a miRNA is defined as follows. The threshold for the portioning is a value to minimize class entropy. TN is the number of training samples.

$$IG(g_i) = \sum_{i=1}^{k} \left(\frac{l_i}{TN}\log\frac{l_i}{n_l} + \frac{r_i}{TN}\log\frac{r_i}{n_r}\right) - \sum_{i=1}^{k} \left(\frac{l_i + r_i}{TN}\right)\log\left(\frac{l_i + r_i}{TN}\right)$$

3.1.3 Mutual Information

Mutual information provides information on the dependency relationship between two probabilistic variables of events. If two events are completely independent, the mutual information is 0. The more they are related, the higher the mutual information is.

$$\begin{aligned} MI(g_i) = {} & MI(g_i \geq \overline{g_i}, t_i = NORMAL) + MI(g_i \geq \overline{g_i}, t_i = TUMOR) \\ & + MI(g_i < \overline{g_i}, t_i = NORMAL) + MI(g_i < \overline{g_i}, t_i = TUMOR) \end{aligned}$$

$$\overline{g}_i = \frac{1}{TN}\sum_{j=1}^{TN} g_{ji}$$

$$MI(g_i > \overline{g_i}, t_i = NORMAL) = P(g_i > \overline{g_i}, t_i = NORMAL)\log_{10}\frac{P(g_i > \overline{g_i}, t_i = NORMAL)}{P(g_i > \overline{g_i}) \times P(t_i = NORMAL)}$$

3.1.4 Signal-to-Noise Ratio

If we calculate the mean μ and standard deviation σ from the distribution of miRNA expressions within their classes, the signal-to-noise ratio (SN) of miRNA g_i is defined as follows:

$$SN(g_i) = \frac{|\mu_{NORMAL}(g_i) - \mu_{TUMOR}(g_i)|}{\sigma_{NORMAL}(g_i) + \sigma_{TUMOR}(g_i)}$$

3.2 Classifiers

3.2.1 K-Nearest Neighbor (KNN)

This is one of the most common methods for instance-based induction. Given an input vector, KNN extracts the *k* closest vectors in the reference set based on similarity measures, and makes a decision for the label of the input vector by using the labels of the *k* nearest neighbors. In this paper, many similarity measures were used such as the inverse of Euclidean distance (KNNE), Pearson correlation (KNNP), cosine coefficients (KNNC) and Spearman correlation (KNNS). If the *k* is not 1, the final outcome is based on the majority voting of the *k* nearest neighbors.

3.2.2 Multi-Layer Perceptron (MLP)

A feed-forward multilayer perceptron is an error backpropagation neural network that can be applied to pattern recognition problems. It requires engineering regarding the architecture of the model (the number of hidden layers, hidden neurons, and so on). In this classification problem, the number of output nodes is two (normal and tumor nodes). If the output from the normal node is larger than that from the tumor node, the sample is classified as normal.

3.2.3 Support Vector Machine (SVM)

This method classifies the data into two classes. SVM builds up a hyperplane as the decision surface in such a way as to maximize the margin of separation between positive and negative samples. In this paper, linear kernel (SVML) is used.

4 Experimental Results

We have used miRNA samples from Lu *et al.*'s work [8]. It contains expression values of 217 miRNAs from 186 samples including multiple cancer types (Table 2). In this work, we did binary classifications which classify samples as one of tumor or normal.

The expression level of each miRNA is normalized to 0~1. For miRNAs, we found the maximum and minimum expression values. The miRNA expression value is adjusted to (g-min)/(max-min). In the feature selection, the number of top-ranked miRNAs is 25. There is no report on the optimal number of miRNAs, but our previous study on mRNA expression profiles indicates that 25 is reasonable [2]. For Information Gain feature selection, we implemented it based on the RANKGENE source code and our IG method showed the same results with the RANKGENE [13]. We used LIBSVM for the SVM classification [14]. The parameters of classification algorithms are summarized in Table 3. The final results are an average of 10 runs. For each run, the miRNA expression data are randomly separated to the training dataset (2/3) and test dataset (1/3).

Table 2. The number of samples for each cancer type

Cancer	Normal	Tumor
Stomach	6	0
Colon	5	10
Pancreas	1	9
Liver	3	0
Kidney	3	5
Bladder	2	7
Prostate	8	6
Ovary	0	7
Uterus	9	10
Human Lung	4	6
Mesothelioma	8	0
Melanoma	0	3
Breast	3	6
Brain	2	0
B Cell ALL	0	26
T Cell ALL	0	18
Follicular Cleaved Lymphoma	0	8
Large B Cell Lymphoma	0	8
Mycosis Fungoidis	0	3
Sum	54	132

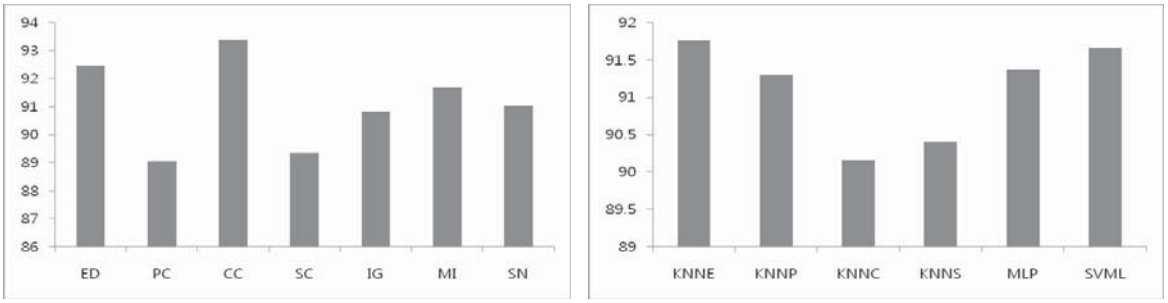
Table 3. Parameters of classification algorithms

Classifier	Parameter	Value
MLP	# of input nodes	25
	# of hidden nodes	8
	# of output nodes	2
	Learning rate	0.05
	Momentum	0.7
	Learning algorithm	Back propagation
KNN	k	3
SVM	Kernel function	Linear

Table 4 shows the comparison of accuracy on test data for the 42 combinations of feature selection and classifications. It shows that the KNNS-CC combination is the best accuracy 95% among them. Figure 2 shows the comparison of average performance of feature selection and classification methods. In the feature selection methods, CC is the best one. However, in the classification algorithm, KNNE is the best one. This means that it is important to find the appropriate combination of feature selection and classification algorithm to get the best accuracy from the miRNA expression profiles. Table 5 shows relevant miRNAs selected by CC methods.

Table 4. Accuracy on test data

	ED	PC	CC	SC	IG	MI	SN
KNNE	92.7	92.4	91.7	90.0	91.7	93.0	90.8
KNNP	93.3	86.4	94.1	87.5	92.9	91.7	93.2
KNNC	92.2	85.9	93.2	87.7	90.6	90.6	90.9
KNNS	93.0	85.3	**95.0**	87.9	89.5	90.9	91.2
MLP	92.0	92.5	94.0	91.2	89.3	91.1	89.5
SVML	91.6	91.7	92.2	91.7	90.9	92.9	90.6



(a) Feature Selection (b) Classification Algorithms

Fig. 2. Comparison of average performance of feature selection and classification methods

Table 5. Relevant miRNAs selected by cosine coefficient

Value	Description
0.814328	hsa-miR-146:UGAGAACUGAAUUCCAUGGGUU:bead_109-A
0.812209	hsa-miR-296:AGGGCCCCCCCUCAAUCCUGU:bead_105-C
0.808118	hsa-miR-21:UAGCUUAUCAGACUGAUGUUGA:bead_119-B
0.805954	hsa-let-7a:UGAGGUAGUAGGUUGUAUAGUU:bead_159-B
0.803176	hsa-miR-16:UAGCAGCACGUAAAUAUUGGCG:bead_105-A
0.799869	hsa-let-7c:UGAGGUAGUAGGUUGUAUGGUU:bead_110-A

5 Conclusions

In this paper, we explore the feature selection and classification algorithms for miRNA expression profiles to classify human cancer. Compared to mRNA expression profile, there are few works using machine learning tools for miRNA data. In this work, we applied seven feature selection methods and six classification algorithms to find the best combination of them. Experimental results show that KNNS + CC method records the best accuracy 95%. For feature selection method, cosine coefficient is the best method. For classification algorithm, KNNE is the superior method. In conclusion, it is important to choose the proper combination of feature selection and classification algorithm to get the high accuracy for miRNA expression profiles.

Acknowledgements

This research was supported by Basic Science Research Program and the Original Technology Research Program for Brain Science through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0012876) (2010-0018948).

References

1. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., et al.: Moecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
2. Cho, S.B., Won, H.H.: Machine learning in DNA microarray analysis for cancer classification. In: The First Asia Pacific Bioinformatics Conference (2003)
3. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al.: Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics 34, 166–176 (2003)
4. Stanford Microarray Database, http://smd.stanford.edu/
5. Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/
6. Ambros, V.: The functions of animal microRNAs. Nature 431, 350–355 (2004)
7. Bartel, D.: MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116, 281–297 (2004)
8. Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., et al.: MicroRNA expression profiles classify human cancers. Nature 435, 834–838 (2005)
9. Xu, R., Xu, J., Wunsch II, D.C.: MicroRNA expression profile based cancer classification using Default ARTMAP. Neural Networks 22, 774–780 (2009)
10. Zheng, Y., Kwoh, C.K.: Informative MicroRNA expression patterns for cancer classification. In: Li, J., Yang, Q., Tan, A.-H. (eds.) BioDM 2006. LNCS (LNBI), vol. 3916, pp. 143–154. Springer, Heidelberg (2006)
11. Cho, S.B.: Exploring features and classifiers to classify gene expression profiles of acute leukemia. International Journal of Pattern Recognition and Artificial Intelligence 16(7), 831–844 (2002)
12. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517 (2007)
13. Su, Y., Murali, T.M., Pavlovic, V., Schaffer, M., Kasif, S.: RankGene: Identification of diagnostic genes based on expression data. Bioinformatics 19, 1578–1579 (2003)
14. http://www.csie.ntu.edu.tw/~cjlin/libsvm/