POSITIVE IMPRESSION OF LOW-RANKING MICRORN AS IN HUMAN CANCER CLASSIFICATION

Feifei Li, Yongjun Piao, Meijing Li, Minghao Piao, Keun Ho Ryu

DataBase&Bioinformatics Laboratory, Dept. of Computer Science,
Chungbuk National University, Cheongju, South Korea
{feifeili, pyz, mjlee, bluemhp, khryu}@dblab.chungbuk.ac.kr

ABSTRACT

Recently, many studies based on microRNAs (miRNAs) showed a new aspect of cancer classification, and feature selection methods are used to reduce the high dimensionality of miRNA expression data. These methods just consider the problem of where feature to class is 1:1 or n:1. But one miRNA may have influence to more than one type of cancers. However, these miRNAs are considered to be low ranked in traditional feature selection methods and they are removed at most of time. Therefore, it is necessary to consider the problem of 1:n or m:n during feature selection. In our wok, we considered both high and low-ranking features to cover all problems (1:1, n:1, 1:n, m:n) in cancer classification. After numerous tests, information gain and chi-squared feature selection methods were chosen to select the high and low-ranking features to form the m-to-n feature subset, and LibSVM classifier was used to do the multi-class classification. Our results demonstrate that the m-to-n features make a positive impression of low-ranking microRNAs in cancer classification since they lead to achieve higher classification accuracy compared with the traditional feature selection methods.

KEYWORDS

low-ranking features, feature selection, cancer classification, microRNA

1. INTRODUCTION

Feature selection, as we know, is aimed to remove the redundant and irrelevant features to find a subset of features. Feature selection method involves two aspects: evaluation of a candidate feature subset using some evaluation criterion, and searching through the feature space to choose a minimum subset of features. Usually, the categories of feature selection algorithms can be identified based on their theoretical basis: correlation, distance, information, dependence, consistency and classifier error rate measures.

The correlation-based feature selection method uses some measures like information gain [1], gain ratio, or linear correlation coefficient [2] to find the good features that highly correlated with the class but not highly correlated with other features. Then these features will be relevant to the class concept but not redundant to any of the other relevant features. And the correlation-based feature selection method has been widely used for many kinds of classification analysis. For the

David C. Wyld et al. (Eds) : CCSIT, SIPP, AISC, PDCTA, NLP - 2014
pp. 449–461, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4239

mutual-information-based feature selection, the largest mutual information reflects the largest dependency in the target class, so the top features are often selected. In research of [3], they proposed a minimal-redundancy-maximal-relevance (mRMR) method, and that can use either mutual information, correlation, distance scores to select features, then tested it with different classifiers, i.e., naive Bayes (NB), support vector machine (SVM), and linear discriminant analysis (LDA). Their results showed that the mRMR can improve the classification accuracy for both discrete and continuous data sets and multiple types of classifiers. Consistency-based search approach [4] uses the inconsistency rate to solve the problem that two instances have the same feature values but have different class labels. This measure is aimed to search in the set of features and find a minimal set of features which are consistent. Feature ranking method sorts features based on the criterion measure, and the criterion measure can be the information, the relevance, or the relation of the features.

Recently, these feature selection methods have been used for cancer classification. With the discovery of microRNAs (miRNAs), a class of small non-coding RNAs, which have been proved that the abnormal expression data can indicate human cancer [5, 6], many feature selection and classification methods have been used to do the miRNA expression data analysis for cancer classification. However, from the year 1993 when the first identified miRNA [7] has been discovered until now, only more than one thousand miRNAs have been discovered. One special characteristic of miRNA expression data is the high dimensionality. The high dimensionality may cause a series of problems for cancer classification, such as add noise, reduce the accuracy rate, and increase the complexity. Although we can use both feature selection and feature extraction to reduce the dimensionalities, feature selection is a better choice than feature extraction for miRNA expression data analysis: feature selection is used in the area where there are a large number of features compared with the small number of samples which is just the characteristic of miRNA expression data; the feature extraction is aimed to create new features using some transform functions of the original features, but these new features maybe cannot be explained in the physical aspect.

However, these methods just consider the condition that the relationship between feature and class is 1:1 or n:1, but not consider the condition that the relationship between feature and class is 1:n or m:n. But since the miRNA expression data is a special kind of data, one miRNA may has influence to more than one type of cancers [8]. If using the traditional feature selection algorithms, these miRNAs may be deleted, since they will be considered as the low-ranking features. But this kind of miRNAs are also very important, removing them may lead to the loss of important information. In Lu et al.'s work [9], they used bead-based flow cytometric miRNA expression profiling method to analyze the 217 mammalian miRNAs from 334 samples including human cancers. And the result showed the potential of miRNA profiling in cancer diagnosis. Based on this data resource, many works using different feature selection methods and classification methods have done to do the cancer classification [10-12]. Most of them based on the binary-class classification, and they showed very high accuracy results. However, these work just considered the high-ranking microRNAs. Therefore, in our study, we made a new hypothesis that consider both of the high and low-ranking features to cover all the cases (1:1, n:1, 1:n, m:n) can get better accuracy in the cancer classification. We used the data resource from Lu et al.'s work, also used different kinds of feature ranking methods with different classifiers to do the analysis. Finally, the results proved that the m-to-n features can get higher classification accuracy compared with the traditional feature selection methods, and it is reasonable to take the low-ranking features into consideration for cancer classification.

The remainder paper is organized as follows. The methods used in the work are discussed in Section 2. Section 3 is about the data set and performance evaluation. The conclusion of our work is presented in Section 4.

2. FEATURE SELECTION AND CLASSIFICATION METHODS

Since there is no evidence to show which kind of feature selection and classification method would fit for miRNA expression data, we chose many different kinds of methods to do the analysis and compared their results.

For feature selection, we used the correlation-based feature selector (CFS) with different search algorithms: re-ranking search algorithm [13], best first search algorithm, particle swarm optimization (PSO) search algorithm [14, 15], and tabu search algorithm [16, 17]. We also used the ranker search method with different attribute evaluators: Pearson's correlation, chi-squared distribution, information gain, and gain ratio.

Re-ranking search algorithm first uses a filter measure to rank all the attributes in decreasing order, the ranking is split in many blocks, and then runs a filter-wrapper algorithm over the first block to select some attributes. Then the remaining attributes are re-ranked again. And the filter-wrapper algorithm is run again on the first current block, and so on. The process is iterated until no attribute is selected in current block. The re-ranking search algorithm can reduce the CPU time and wrapper evaluations compared with the incremental feature subset selection algorithms.

Best first search algorithm first evaluates all the features regard them as a separate subset. And the feature subset which has the highest object function is selected. Then the algorithm updates the subset by adding all the possible combinations of new single features. After that the algorithm evaluates the new subset, if the result is improved then the features are retained otherwise the process searches other features to expend the subset. The process will stop when there is no improvement by adding new features.

Particle swarm optimization (PSO) is usually applied to continuous search spaces. It is a population-based heuristic global optimization algorithm for feature selection. In the PSO algorithm, particle swarm involves n particles, each particle has a randomized velocity based an objective function. And the particle iteratively changes its position to find most optimist position of particle itself and swarm. The algorithm stops until a termination criterion is met.

Tabu search is a metaheuristic search method used for solving mathematical optimization problems. It is a form of local neighborhood search. The tabu search algorithm starts with a current solution, and evaluates the criterion function for that solution. Then the algorithm stores the neighboring solutions in the candidate list. And then it finds the best candidate from the candidate list, if the candidate has a higher criterion function value than the current best, its features are add to the tabu list and it is viewed as the new best. The process is looped until the stop criterion is met. At last the best solution obtained so far is the solution of the tabu search.

The Pearson's correlation coefficient is used to compute the correlation between the feature and the class variable. It is defined by the following equation (1):

$$\mathrm{PCC} = \frac{\sum(x_i - \bar{x})(c_i - \bar{c})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(c_i - \bar{c})^2}} \qquad (1)$$

where the index ci is the class label. The Pearson's correlation coefficient ranges from -1 and 1. If the value is 1 (or -1) that means the feature and class label have a perfect positive (or negative)

linear relationship. However, if the value is 0, then there is no linear relationship between the feature and class label.

Chi-squared attribute evaluation evaluates a feature by computing the chi-squared statistic of the feature with respect to the class label. First the hypothesis H0 is assumed as the two features are unrelated. Then it is tested using the following equation (2):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (2)$$

where Oij is the observed frequency and Eij is the expected frequency, the larger the value of $\chi 2$ is, the more evidence to show that the hypothesis H0 is true.

Information gain measures the expected reduction in entropy. The entropy is a measure of the uncertainty in a random variable. Based on it, the information gain of a feature is defined as the following equation (3):

$$\Delta_{info} = \text{Entropy}(C) - \sum_{j=1}^{k} \frac{|C_v|}{|C|} \text{Entropy}(C_v) \quad (3)$$

where k is the number of attribute values, C is a collection of samples and Cv is the subset of collection C for attribute which has the value of v. The higher value indicates the higher purity of class.

Gain ratio is aim to maximize the information gain of feature and minimize the number of its value. Gain ratio is the ratio between the information gain and intrinsic value defined as the following equation (4):

$$\text{Gain Ratio} = \frac{\Delta_{info}}{-\sum_{j=1}^{k} \frac{|C_v|}{|C|} \log_2 \frac{|C_v|}{|C|}} \quad (4)$$

For classification methods, we chose four different kinds of classifiers: support vector machine (SVM), naive Bayes, k-nearest neighbors (KNN), and decision tree. The SVM constructs a hyperplane or sets of hyperplanes in a high dimensional space, and aims to find the largest margin to separate the objects of different classes. To build the SVM classifier, we used the LibSVM package [18] since it can support both 2-class and multi-class classification. A naive Bayes classifier assumes that each feature is independent to others. And a naive classifier is based on the Bayes' theorem and is very efficiently in supervised learning. For the naive Bayes classifier, we adopted the Aggregating One-Dependence Estimators (A1DE) algorithm [19]. This algorithm can solve the attribute-independence problem of the traditional naive Bayed classifier. The k-nearest neighbors classifier can predict the class label of the object based on the k closest objects in the feature space. The KNN algorithm is almost the simplest of all machine learning algorithms. Decision tree classifier is a form of a tree structure. In the tree, each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The path from the root to the leaf represents classification rules. And we chose the

C4.5 algorithm to build the decision tree classifier. This algorithm uses information gain as the splitting criterion.

3. EXPERIMENTAL RESULTS

3.1. Data Set

The miRNA expression data used in this paper is from Lu et al.'s work. It is used to build a multi-class classifier, it consists of five kinds of tumor samples from colon, uterus, pancreas, T-cell ALL, and B-cell ALL which totally includes 73 samples with the expression value of 217 miRNAs for multiple cancer types. The detail of the cancer types shows in Table 1.

Table 1. The number of the samples for each cancer type

Cancer Name	Number of Tumor Samples
Colon	10
Pancreas	9
Uterus	10
B Cell ALL	26
T Cell ALL	18
SUM	73

3.2. Performance Evaluation

To get a reliable result, 10-fold cross validation is performed on the whole data set. And the data set is randomly divided into 10 parts, nine of them are used as training set, the rest part is used as test set.

In our study, we first used the correlation-based feature subset selection methods with four different search methods: re-ranking search, best first search, tabu search and PSO search method. Using these search methods can automatically select the features with the exactly number. For comparison, we tested these features on four classifiers including LibSVM algorithm of SVM classifier, A1DE algorithm of naive Bayes classifier, J48 algorithm of decision tree classifier and IBK algorithm of k-nearest neighbor classifier. Table 2 shows the final results, after feature selection, the re-ranking search method resulted in 15 top-ranking features, the best search method resulted in 16 top-ranking features, the tabu search method resulted in 17 top-ranking methods, and the PSO search method resulted in 50 top-ranking features. The LibSVM algorithm shows the better results with the accuracy of 91.78% without feature selection. However, after we reduced the dimensionality of miRNA expression data, the result became quite different. With the feature selection, most of the accuracies of the classification methods have been increased except LibSVM classifier. The A1DE classifier got better result when using the re-ranking, best first, and tabu search method. And when using the PSO search method, the results of J48 and IBK classifiers have been increased. The result indicated that feature selection is very necessary for cancer classification. However, these methods just selected the fit number of features, since the number is very small, we cannot find how the feature number influences the classification accuracy. Therefore, we need to use other methods to find the relationship between feature number and classification accuracy.

Table 2. Classification accuracy (%) of four classification algorithms (FS: feature selection).

Method	LibSVM	A1DE	J48	IBK
Without FS	91.78	86.30	86.30	83.56
Re-ranking	90.41	89.04	84.93	83.56
Best First	90.41	89.04	84.93	83.56
Tabu	90.41	90.41	84.93	82.91
PSO	89.04	84.93	87.67	86.30

Classification algorithms: LibSVM, A1DE, J48 and IBK. Feature selection method: correlation-based subset selection algorithm with different search methods including re-ranking search, best first search, tabu search and PSO search method. The selected high-ranking feature number with these four methods is : 15, 16, 17 and 50.

In order to find the relationship, we did another experiment used the Pearson's correlation, chi-squared distribution, information gain, and gain ratio as the attribute evaluators to do the feature selection. The LibSVM package solved the quadratic problems and used shrinking and caching methods to reduce the size of working problem, also in the last experimental results the LibSVM got better accuracy compared with other classification methods. Therefore, the LibSVM package was chosen as the classifier. Figure 1, 2, 3, 4 shows the classification accuracy for the four kinds of feature selection methods (i.e. Pearson's correlation, chi-squared, gain ratio, information gain) with LibSVM classifier. The top-ranking feature number that we chose for test is from 10 to 210. Compared both the results of the four feature selection methods, the Pearson' correlation method and gain ratio method show the similar results, and the chi-squared method and information gain method show the similar results. For Pearson's correlation method and gain ratio method, when the feature number is very small, the accuracy is very low, but the accuracy of chi-squared method and information gain method is high. For both of these feature selection methods, there is a same trend that with the increase of the feature numbers the accuracy is also been improved. Also figure 5 is the histogram of these four feature selection methods. When the feature number is smaller than 20, the chi-squared and information gain methods show higher accuracy comparing with the Pearson's correlation and gain ratio methods. When the feature number is between 50 and 140, the Pearson's correlation and gain ratio methods show the better result. When the feature number is larger than 170, these methods show the similar high accuracy.

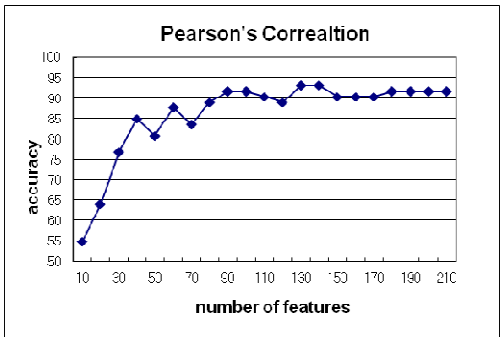

Figure 1. Classification accuracy (%) for the Pearson's correlation feature selection method with LibSVM classifier. The number of the selected high-ranking features is from 10 to 210.

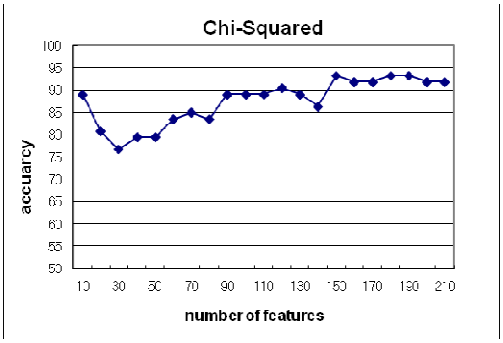

Figure 2. Classification accuracy (%) for the chi-squared feature selection method with LibSVM classifier. The number of the selected high-ranking features is from 10 to 210.

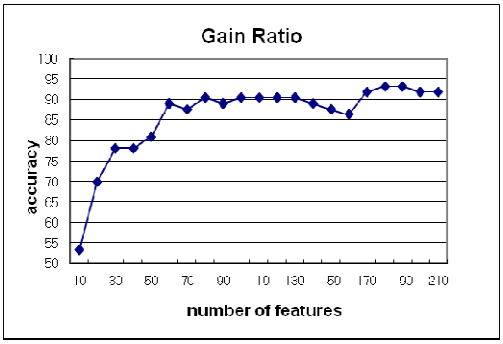

Figure 3. Classification accuracy for gain ratio feature selection method with LibSVM classifier. The number of the selected high-ranking features is from 10 to 210.

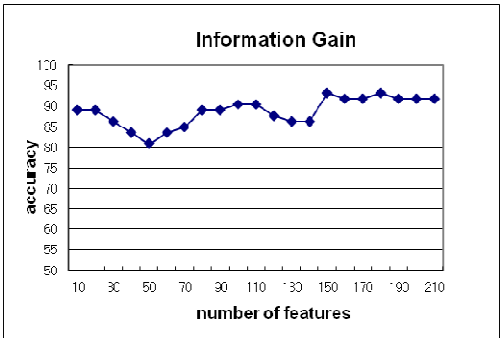

Figure 4. Classification accuracy for information gain feature selection method with LibSVM classifier. The number of the selected high-ranking features is from 10 to 210.

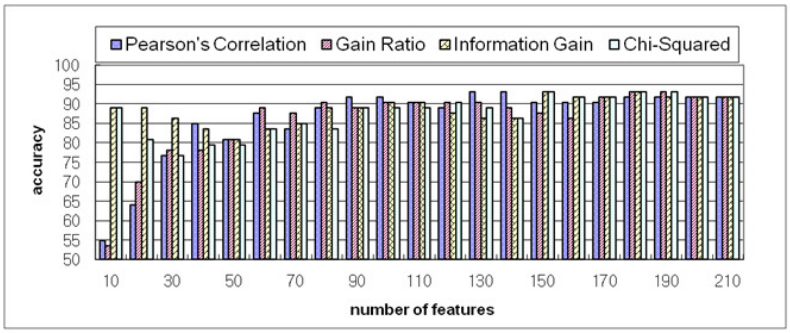

Figure 5. The histogram of the four feature selection methods (i.e. Pearson's correlation, gain ratio, information gain, chi-squared) with LibSVM classifier. The number of the selected high-ranking features is from 10 to 210.

Both of these feature selection methods select the high-ranking features, if we want to get the higher accuracy, the feature number should be large enough, but the large number is meaningless. Therefore, we considered both the high and low-ranking features to form the m-to-n feature subset. The previous experiment shows that the information gain and chi-squared feature selection methods are totally better compared with the other two methods. Because when the feature number is small, the Pearson's correlation and gain ratio feature selection methods show very low classification accuracy which means these selected top-ranking features cannot excellent classify the miRNA data. Considering this reason, the information gain and chi-squared feature selection methods were used to form the feature subsets with both the high and low-ranking features, and as well the LibSVM package of SVM classifier was chosen for the multiple classification problems.

The result is showed in Table 3. First we chose 10 high-ranking features, and that means the relationship between feature and class is 1 to 1 or n to 1. The information of selected high-ranking microRNA is showed in table Table 4 and Table 5. The classification accuracy is 89.04% for both of the two feature selection methods. Then we considered the case of the feature to class is 1:n, in this case we selected 17 low-ranking features. The information of selected low-ranking microRNA is showed in Table 6 and Table 7. The classification accuracy of information gain method is 52.05% while the classification accuracy of chi-squared method is 50.68%. Obviously the accuracy is very low since the low-ranking features would lead to the impurity of the class. At last, we considered the m-to-n features with both the high-ranking and low-ranking features, and in this condition feature to class is m:n. We combined both the 10 high-ranking features and 17 low-ranking features together, totally 27 features, and used them to do the classification, surprisingly, we got a very good result, with the classification accuracy of information gain method is 94.52% and the classification accuracy of chi-squared method is 93.14%. In the work of [11], they used the Default ARTMAP as the classifier to do the multi-class cancer classification with the same data set as in our work. But the best result only has the accuracy of 88.89%. Compared with this work and our first experiment in Table 2, feature selection with the m-to-n features got the highest classification accuracy. The result also proved that it is reasonable to take the low-ranking features into consideration when doing cancer classification.

Table 3. Classification accuracy for LibSVM classifier considering the high-ranking and low-ranking features. 1:1, n:1, 1:n and m:n indicate the relationship between feature and class.

Relationship	Information Gain	Chi-Squared
1:1 or n:1	89.04	89.04
1:n	52.05	50.68
m:n	94.52	93.14

Table 4. The information of 10 high-ranking microRNA selected by Information Gain method.

Probe ID	Target Sequence	MicroRNA Name
EAM250	AUGACCUAUGAAUUGACAGAC	hsa-miR-215
EAM330	UGUAAACAUCCUCGACUGGAAGC	hsa-miR-30a-5p
EAM105	UCCCUGAGACCCUAACUUGUGA	hsa-miR-125b
EAM348	CAUCAAAGUGGAGGCCCUCUCU	mmu-miR-291-5p
EAM190	UACCCUGUAGAACCGAAUUUGU	hsa-miR-10b
EAM288	CCCUGUAGAACCGAAUUUGUGU	mmu-miR-10b
EAM366	UUCAGCUCCUAUAUGAUGCCUUU	mmu-miR-337
EAM261	AUCACAUUGCCAGGGAUUACCAC	hsa-miR-23b
EAM260	AUCACAUUGCCAGGGAUUUCC	hsa-miR-23a
EAM381	UCGAGGAGCUCACAGUCUAGUA	rno-miR-151*

Table 5. The information of 10 high-ranking microRNA selected by Chi-Squared method.

Probe ID	Target Sequence	MicroRNA Name
EAM250	AUGACCUAUGAAUUGACAGAC	hsa-miR-215
EAM190	UACCCUGUAGAACCGAAUUUGU	hsa-miR-10b
EAM288	CCCUGUAGAACCGAAUUUGUGU	mmu-miR-10b
EAM105	UCCCUGAGACCCUAACUUGUGA	hsa-miR-125b
EAM366	UUCAGCUCCUAUAUGAUGCCUUU	mmu-miR-337
EAM381	UCGAGGAGCUCACAGUCUAGUA	rno-miR-151*
EAM303	UACAGUAGUCUGCACAUUGGUU	hsa-miR-199a*
EAM336	AGGCAGUGUAGUUAGCUGAUUGC	hsa-miR-34c
EAM339	CACCCGUAGAACCGACCUUGCG	hsa-miR-99b
EAM260	AUCACAUUGCCAGGGAUUUCC	hsa-miR-23a

Table 6. The information of 17 low-ranking microRNA selected by Information Gain method.

Probe ID	Target Sequence	MicroRNA Name
EAM247	UAACAGUCUCCAGUCACGGCC	hsa-miR-212
EAM252	UACUGCAUCAGGAACUGAUUGGAU	hsa-miR-217
EAM254	UGAUUGUCCAAACGCAAUUCU	hsa-miR-219
EAM259	UGUCAGUUUGUCAAAUACCCC	hsa-miR-223
EAM283	UUCCCUUUGUCAUCCUUUGCCU	mmu-miR-211
EAM293	CAUCCCUUGCAUGGUGGAGGGU	hsa-miR-188
EAM306	UACUCAGUAAGGCAUUGUUCU	mmu-miR-201
EAM308	UGGAAUGUAAGGAAGUGUGUGG	hsa-miR-206
EAM309	GCUUCUCCUGGCUCUCCUCCCUC	mmu-miR-207
EAM328	CAGUGCAAUAGUAUUGUCAAAGC	hsa-miR-301
EAM331	UGUAAACAUCCUUGACUGGA	hsa-miR-30e
EAM337	CAAAGUGCUGUUCGUGCAGGUAG	hsa-miR-93
EAM340	CUAUACGACCUGCUGCCUUUCU	mmu-let-7d*
EAM341	CAAAGUGCUAACAGUGCAGGUA	mmu-miR-106a
EAM346	CUCAAACUAUGGGGGCACUUUUU	mmu-miR-290
EAM352	AAAGUGCUUCCCUUUUGUGUGU	mmu-miR-294
EAM361	CCUCUGGGCCCUUCCUCCAGU	hsa-miR-326

Table 7. The information of 17 low-ranking microRNA selected by Chi-Squared method.

Probe ID	Target Sequence	MicroRNA Name
EAM247	UAACAGUCUCCAGUCACGGCC	hsa-miR-212
EAM252	UACUGCAUCAGGAACUGAUUGGAU	hsa-miR-217
EAM254	UGAUUGUCCAAACGCAAUUCU	hsa-miR-219
EAM259	UGUCAGUUUGUCAAAUACCCC	hsa-miR-223
EAM283	UUCCCUUUGUCAUCCUUUGCCU	mmu-miR-211
EAM290	UGGACGGAGAACUGAUAAGGGU	hsa-miR-184
EAM293	CAUCCCUUGCAUGGUGGAGGGU	hsa-miR-188
EAM308	UGGAAUGUAAGGAAGUGUGUGG	hsa-miR-206
EAM309	GCUUCUCCUGGCUCUCCUCCCUC	mmu-miR-207
EAM324	CAUUGCACUUGUCUCGGUCUGA	hsa-miR-25
EAM328	CAGUGCAAUAGUAUUGUCAAAGC	hsa-miR-301
EAM331	UGUAAACAUCCUUGACUGGA	hsa-miR-30e
EAM337	CAAAGUGCUGUUCGUGCAGGUAG	hsa-miR-93
EAM340	CUAUACGACCUGCUGCCUUUCU	mmu-let-7d*
EAM341	CAAAGUGCUAACAGUGCAGGUA	mmu-miR-106a
EAM346	CUCAAACUAUGGGGGCACUUUUU	mmu-miR-290
EAM352	AAAGUGCUUCCCUUUUGUGUGU	mmu-miR-294

4. CONCLUSIONS

The right choice of feature selection and classification method is very important to cancer classification since the special characteristic of miRNA expression data. After numerous tests, the information gain and chi-squared feature selection methods were chosen to do the dimensionality reduction. Different with the traditional feature selection, we considered all cases (1:1, n:1, 1:n, m:n) in cancer classification. Our work has proved the usefulness of the m-to-n features in cancer classification, since the results showed that considering both the high-ranking and low-ranking features can get higher classification accuracy than just considering the high-ranking features. And the selected low-ranking miRNAs in Table 6 and Table 7 provide cancer researchers some very useful information for further research analysis of their function in human cancer. However, there have some shortcomings: we have tested for many times to find a relatively good number of the m-to-n features to do the analysis, but in fact it is very difficult to determine the best number of the selected features.

In the future work, we will do our best to discover some feature selection algorithms which can choose the appropriate m-to-n feature number automatically. Also we will try to use this idea to test for other kinds of data not only the miRNA expression data.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2008-0062611).

REFERENCES

[1] Lei Yu, Huan Liu: Feature selection for high-dimensional data: a fast correlation-based filter solution. AAAI. 856-863 (2003).
[2] Rajdev Tiwari, Manu Pratap Singh: correlation-based attribute selection using genetic algorithm. IJCA. 4, 28-34 (2010).
[3] Hanchuan Peng, Fuhui Long, Chris Ding: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE. 27, 1226-1238 (2005).
[4] Manoranjan Dash, Huan Liu: Consistency-based search in feature selection. Artificial Intelligence. 151, 155-176 (2003).
[5] He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ, Hammond SM: A microRNA polycistron as a potential human oncogene. Nature. 435, 828-833 (2005).
[6] Mraz M, Pospisilova S, Malinova K, Slapak I, Mayer J: MicroRNAs in chronic lymphocytic leukemia pathogenesis and disease subtypes. Leuk Lymphoma. 50, 506-509 (2009).
[7] Lee RC, Feinbaum RL, Ambros V: The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 75, 843-854 (1993).
[8] Moller HG, Rasmussen AP, Andersen HH, Johnsen KB, Henriksen M, Duroux M: A systematic review of microRNA in glioblastoma multiforme: micro-modulators in the mesenchymal mode of migration and invasion. Mol Neurobiol. 47, 131-144 (2013).
[9] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: MicroRNA expression profiles classify human cancers. Nature. 435, 834-838 (2005).

[10] Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, Levy A, Lebanony D, Goren Y, Silberschein E, Targan N, Ben-Ari A, Gilad S, Sion-Vardy N, Tobar A, Feinmesser M, Kharenko O, Nativ O, Nass D, Perelman M, Yosepovich A, Shalmon B, Polak-Charcon S, Fridman E, Avniel A, Bentwich I, Bentwich Z, Cohen D, Chajut A, Barshack I: MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol. 26, 462-469 (2008).
[11] Xu R, Xu J, Wunsch DC 2nd: MicroRNA expression profile based cancer classification using Default ARTMAP. Neural Networks. 22, 774-780 (2009).
[12] Kyung-Joong Kim, Sung-Bae Cho: Exploring features and classifiers to classify microRNA expression profiles of human cancer. Neural Information Processing. 6444, 234-241 (2010).
[13] Pablo Bermejo, Luis de la Ossa, Jose A. Gamez, Jose M. Puerta: Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. Knowledge-Based Systems. 25, 35-44 (2012).
[14] Alberto Moraglio, Cecilia Di Chio, Julian Togelius, Riccardo Poli: Geometric particle swarm optimization. Journal of Artificial Evolution and Application. 2008, 14 pages (2008).
[15] J.Garcia-Nieto, E.Alba, L.Jourdan, E.Talbi: Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. Information Processing Letters. 109, 887-896 (2009).
[16] Abdel-Rahman Hedar, Jue Wang, Masao Fukushima: Tabu search for attribute reduction in rough set theory. Soft Computing. 12, 909-918 (2008).
[17] Hongbin Zhang, Guangyu Sun: Feature selection using tabu search method. Pattern Recognition. 35, 701-711 (2002).
[18] Chih-Chung Chang, Chih-Jen Lin: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent System and Technology. 2 (2011).
[19] Geoffrey I.Webb, Janice R.Boughton, Zhihai Wang: Not So Naive Bayes: Aggregating One-Dependence Estimators. Machine Learning. 58, 5-24 (2005).

AUTHORS

Feifei Li was born in Shandong Province, China, in 1989. She received an Undergraduate degree in 2012 from Harbin Institute of Technology, China. Now she is pursuing the Master degree at Electrical and Computer Engineering, Chungbuk National University, South Korea. Her research fields include data mining application to human cancer, classification analysis of high-dimensional data, and feature reduction for microRNA expression data.

Yongjun Piao was born in Jilin Province, China, in 1987. He received a Master degree in Computer Science in 2013 from Chungbuk National University, South Korea. Currently, he is pursuing the Ph.D. degree at the Computer Science, Chungbuk National University, South Korea. His research activities include bioinformatics, computational biology, data mining.

Meijing Li received a MS degree at Database and Bioinformatics Laboratory, Chungbuk National University, Cheongju, South Korea in 2010. She received BS degree in the School of Information and Computing Science from Dalian University, China, in 2007. Currently, she is a Ph.D. candidate at the same laboratory of the Department of Computer Science, Chungbuk National Univ., Rep. of Korea since 2010. Her major research interests include database, bioinformatics and data mining.

Minghao Piao was born in Jilin Province, China, in 1983. He received a Master degree in Bioinformatics in 2009 from Chungbuk National University, South Korea. Currently, he is pursuing the Ph.D. degree at the Computer Science, Chungbuk National University, South Korea. He became a member of IEEE in April, 2012.His research activities include electrical customer classificati on, data mining application to energy data, discovery of emerging patterns in clustering analysis, and reduction of information redundancy for building classifiers.

Keun Ho Ryu received the Ph.D. degree from Yonsei University, Seoul, Korea, in 1988. He is a Professor in the School of Electrical and Computer Engineering, the Chungbuk National University, Cheongju, Korea. He was a Postdoctoral Researcher at the University of Arizona and also a Research Scientist at Electronics and Telecommunications Research Institute, Daejeon,Korea. His research interests include temporal databases, spatiotemporal databases, temporal GIS, ubiquitous computing and stream data processing, knowledgebase information retrieval, database security, data mining, and bioinformatics. Dr. Ryu is a member of the IEEE and ACM since 1983.