Computers in Biology and Medicine 80 (2017) 39–44

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles

Yongjun Piao[a], Minghao Piao[b], Keun Ho Ryu[a,*]

[a] Database/Bioinformatics Laboratory, College of Electrical & Computer Engineering, Chungbuk National University, Cheongju, 28644, South Korea
[b] Department of Computer Engineering, Dongguk University Gyoengju Campus, 38066, South Korea

ARTICLE INFO

Keywords:
Cancer classification
Data mining
Ensemble learning
miRNA expression

ABSTRACT

Cancer classification has been a crucial topic of research in cancer treatment. In the last decade, messenger RNA (mRNA) expression profiles have been widely used to classify different types of cancers. With the discovery of a new class of small non-coding RNAs; known as microRNAs (miRNAs), various studies have shown that the expression patterns of miRNA can also accurately classify human cancers. Therefore, there is a great demand for the development of machine learning approaches to accurately classify various types of cancers using miRNA expression data. In this article, we propose a feature subset-based ensemble method in which each model is learned from a different projection of the original feature space to classify multiple cancers. In our method, the feature relevance and redundancy are considered to generate multiple feature subsets, the base classifiers are learned from each independent miRNA subset, and the average posterior probability is used to combine the base classifiers. To test the performance of our method, we used bead-based and sequence-based miRNA expression datasets and conducted 10-fold and leave-one-out cross validations. The experimental results show that the proposed method yields good results and has higher prediction accuracy than popular ensemble methods. The Java program and source code of the proposed method and the datasets in the experiments are freely available at https://sourceforge.net/projects/mirna-ensemble/.

1. Introduction

Cancer is a class of complex genetic diseases that are characterized by out-of-control cell growth. Cancer classification has been a crucial topic of research in cancer treatment. In the last decade, mRNA expression data have been widely used to classify different types of human cancers [1]. Various machine learning approaches have been developed [2–5] to reduce the dimensionality of mRNA expression data and improve the classification accuracy.

With the discovery of a class of small non-coding RNAs, known as microRNAs (miRNAs), the expression patterns of these molecules have attracted the attention of many researchers. miRNAs play important regulatory roles in biological processes such as development, cell proliferation, differentiation and apoptosis [6,7] by pairing the mRNA of protein-coding genes with the transcriptional or post-transcriptional regulation of their expression [8]. miRNAs have emerged as highly tissue-specific biomarkers that function as tumor suppressors and oncogenes. Furthermore, several studies have shown that the expression patterns of miRNAs are heterogeneous in different human cancers [9–11]. Therefore, there is a great demand for developing machine learning approaches to accurately classify various types of cancers from miRNA expression data. Moreover, next-generation sequencing technology is increasingly used to quantify miRNA expression levels (miRNA-seq) as an alternative to microarrays. The classification model should also be scalable to such types of 'digital' expression data.

It is well known that ensembles of classifiers can improve the prediction accuracy by constructing a set of base classifiers from the training data and performing classification by combining the results of each base classifier. Several methods to construct an ensemble have been developed [12,13], such as instance subset-based approaches (i.e. bagging and boosting) and feature subset-based approaches (i.e. random forests). However, there are some difficulties in using the instance subset-based approach for miRNA expression data classification because the main characteristic of the data is that they lack training samples compared with dimensions. The basic idea of a feature subset-based ensemble is simply to give each classifier a different projection of the training set [14]. A feature subset-based ensemble has several advantages: i) automatically removes irrelevant and redundant features, ii) it performs fast because of the reduced size of the input

* Corresponding author.
E-mail addresses: pyz@dblab.chungbuk.ac.kr (Y. Piao), myunghopark@gmail.com (M. Piao), khryu@dblab.chungbuk.ac.kr (K.H. Ryu).

http://dx.doi.org/10.1016/j.compbiomed.2016.11.008
Received 9 September 2016; Received in revised form 15 November 2016; Accepted 20 November 2016

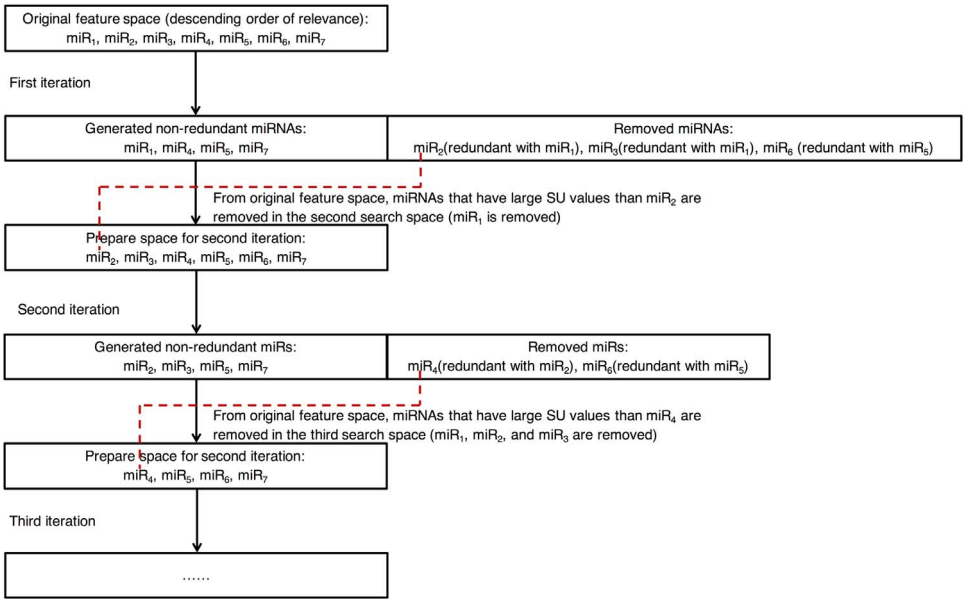


Fig. 1. Example of the generation of multiple miRNA subsets.

space, and iii) it reduces the correlations among the classifiers. Generating feature subsets for an ensemble can be viewed as multiple runs of feature selection procedures. In the past, several feature selection techniques have been proposed, such as information gain, gain ratio, and correlation coefficients. However, these methods are designed to identify a best single-feature subset, which is not suitable for direct application to ensemble generation. Moreover, most of them do not consider the interactions among the features.

In this article, we propose a feature subset-based ensemble method to classify multiple tissues with miRNA expression data. First, we suggest an miRNA subset generation method based on the relevance of miRNAs to cancers and the interactions among miRNAs. Then, a number of classifiers learn from the generated subsets. Finally, the results of each classifier are combined with the average probability of the classifiers. It is observed that the proposed ensemble method obtains promising classification accuracy compared with other ensemble methods.

2. Method

An ensemble method with multiple independent feature subsets is proposed to classify various cancer types. Note that from the data mining viewpoint, an miRNA can be considered as a feature for classification purposes. The proposed method has three major steps: i) generation of multiple miRNA subsets based on the correlations among the miRNAs; ii) learning of the model from each miRNA subset using a machine learning algorithm as a base classifier; iii) combination of the results of each classifier by averaging the probabilities. In the next sections, we will illustrate each step in detail.

2.1. Generation of multiple miRNA subsets

There is no doubt that the major factor to achieve better ensemble performance is the diversity of each ensemble member. The basic idea of our subset generation method is obtaining the diversity by identifying relevant features and putting redundant features into different base classifiers. Various studies have demonstrated that the symmetrical uncertainty (SU) is a good measure to identify both relevant and redundant features [15–17]. The SU is a correlation measure of a random variable based on the information-theoretical concept of entropy. The SU between any pair of features or a feature and class can be calculated as follows:

$$\mathrm{IG}(X|Y) = H(X) - H(X|Y) \quad (1)$$

$$\mathrm{SU}(X, Y) = 2*\mathrm{IG}(X|Y)/(H(X) + H(Y)) \quad (2)$$

where IG(X|Y) is the information gain of X after observing variable Y, and H(X) and H(Y) are the entropy values of variables X and Y, respectively. The SU value is 0–1, where 1 indicates complete correlation and 0 indicates no correlation. To the best of our knowledge, FCBF [18] was the first method to define feature relevance and redundancy using SU.

Definition 1. (Relevant Feature) A feature X is relevant if the SU value to the class, which is denoted as SU(X,C), is larger than a user-defined threshold.

Definition 2. (Redundant Feature) Relevant features X and Y are redundant if their SU, which is denoted as SU(X,Y), is larger than min(SU(X,C), SU(Y,C)).

Given a discretized miRNA expression dataset D with m samples and n miRNAs (miR_1, miR_2..., miR_n), the proposed method first searches all relevant miRNAs and sorts them in descending order of relevance. Note that the irrelevant miRNAs will no longer be considered. Then an miRNA miR_t is selected as the starting point to generate a subset. Similar to FCBF, the proposed method finds all redundant miRNAs and forms a non-redundant feature subset by removing the redundant miRNAs with miR_t. The next key point is how to form the input space to generate the next non-redundant subset. Between two redundant miRNAs, the less relevant miRNA may also produce a competitive result when the combinations of the miRNAs are considered. Therefore, we use the less relevant miRNA as a starting point for the next search. For ease of understanding, let us consider a subset search procedure with 7 miRNAs, as illustrated in Fig. 1. Suppose that all miRNAs are already sorted in descending order of their relevance; then, the most relevant miRNA, miR_1, is selected as a starting point to generate the first subset. Assume that miR_2 and miR_3 are redundant with miR_1, and miR_6 is redundant with miR_5; then, miR_2, miR_3, and miR_6 are removed. The remaining subset, which includes miR_1, miR_4, miR_5, and miR_7, is the generated subset in the first search. In the second iteration, we select the most relevant miRNA among the removed miRNAs (in this case, miR_2) as the starting point. In addition, it is not necessary to analyze the entire feature space again because the

starting point was eliminated by miRNAs with larger SU values than the updated starting point. Thus, we prepare the search space by removing from the original space the miRNAs with larger SU values than the starting point. In this example, miR_1 is excluded from the search space for the second iteration because miR_1 is more relevant than miR_2. Consequently, the subset that contains miR_2, miR_3, miR_4, miR_5, miR_6, and miR_7 is the search space for the second iteration. Then, the redundant miRNAs are removed again, as in the first search. Suppose that miR_4 is redundant with miR_2, and miR_6 is redundant with miR_5; the generated subset in the second iteration is {miR_2, miR_3, miR_5, and miR_7}. From the first iteration, we observe that miR_2 and miR_3 are removed because of redundancy with miR_1, although they have higher relevance values than the others. Obviously, miR_2 and miR_3 appear in the second generated subset by removing miR_1 in the preparation of the second search space. This process may be efficient because we cannot guarantee that miR_1 always makes a better prediction than the combination of miR_2 and miR_3. The subset generation procedure is repeated until we select the user-defined number of subsets.

2.2. Ensemble learning

To construct each base classifier on the generated subsets from step 1, we used a C4.5 decision tree algorithm [19] and a Support Vector Machine (SVM) [20]. We selected these two models because of the common use in data mining applications. Any other classification algorithms can be used as the base classifier for our ensemble method. After constructing a set of classifiers, although we can use one classifier that shows the best performance on the target, some studies mentioned that it might be a waste to use only one classifier committee and ignore the useful information in the other classifiers. To maximize the advantages of the ensemble, the outputs of the classifiers should combine for the final classification. There are several classifier combination methods, such as minimum, maximum, average, median and majority vote of the posteriori probability. Kuncheva [21] shows that these methods have similar performance for the normally distributed classification error, and the average method outperforms the minimum/maximum method for the uniformly distributed error. Therefore, the average of the posteriori probability is used as the combination method in our proposed model.

Consider a classification problem in which object x is to be assigned to one of i possible classes (c_1, ..., c_i). The posterior probability indicates the probability that a given object x belongs to one of the classes. The idea of the average method is to sum up the probabilities for a given instance from each classifier [22] and then obtain the average by dividing the sum by the number of base classifiers.

3. Results

3.1. Datasets

The first expression dataset D_1 was constructed from The Cancer Genome Atlas (TCGA) data portal (https://gdc-portal.nci.nih.gov). First, we collected the raw expression count data of four cancers: 87 breast invasive carcinoma (BRCA) samples, 27 lymphoid neoplasm diffuse large B-cell lymphoma (DLBC) samples, 34 pancreatic adenocarcinoma (PAAD) samples, and 67 prostate adenocarcinoma (PRAD) samples with 1,047 miRNAs. The TCGA data in this study were downloaded on May 21, 2014, and the details of the datasets can be found in Supplementary File 1. Then, we merged these four types together to form D_1. The second expression dataset was first published by Lu et al. [9]. They used a bead-based method to present a systemic expression analysis of 217 mammalian miRNAs from 186 samples, which included multiple human cancers. Based on the original dataset, we constructed D_2 by removing the tumors that included fewer than 9 samples. The details of the datasets are shown in Table 1.

Table 1
Datasets for the experiment.

Dataset	Diseases	Samples	miRNAs
D_1	BRCA	87	1,047
	DLBC	27	
	PAAD	34	
	PRAD	67	
D_2	COLON	10	217
	PAN	9	
	UT	10	
	T_BALL	26	
	T_TALL	18	

3.2. Experimental setup

In our experiments, we selected three widely used ensemble methods to compare with our proposed method. One method is the tree-based ensemble method, random forest [23], which randomly partitions the original feature space and constructs the ensembles by merging base classifiers that are learned from each random subset. The other two methods are the instance-based ensemble algorithm, which iteratively performs instance sampling based on the un-weighted and weighted manners and are denoted as bagging [24] and boosting [25], respectively. In addition, we selected two popular classification methods as the base classifiers of the ensembles except random forest to evaluate the classification performance: C4.5 and support vector machine. Note that the random forest contains a classification algorithm. The base classifiers and other ensemble methods use the WEKA implementation [26]. To solve multi-class problems in the SVM, we used the one-against-one approach, which constructs N(N-1)/2 binary classifiers for an N class dataset. The pairwise coupling method was used to combine the posterior probabilities of individual binary classifiers. For each miRNA dataset, we performed 10-fold cross validation and leave-one-out cross validation. In the 10-fold cross validation, the original data were partitioned into ten equal components; nine were used for training, and the remaining component was used for testing. The leave-one-out cross validation involved using one instance for testing and the remaining instances for training until all instances were tested once. The process was repeated 50 times and the results of each method were recorded.

3.3. Evaluation measures

To quantify the performance of our ensemble method and compare it with existing methods, the classification accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) [27] were used as the predictive measures.

- Accuracy=(no. true positives+no. true negatives)/total no. instances
- Sensitivity=no. true positives/(no. true positives+no. false negatives)
- Specificity=no. true negatives/(no. true negatives+no. false positives)
- AUC is an index used to summarize ROC curve that ranges from 0 to 1

where true positives (TPs) refer to the correct prediction of positive instances, false negatives (FNs) are the incorrect rejection of positive instances, true negatives (TNs) are the correct rejection of negative instances, and false positives (FPs) are the incorrect prediction of negative instances. The performance measures of class i were estimated by considering the i-labeled instances as positives and the remainder as negatives.

Table 2a
Classification results on D_1 (C4.5 as the base classifier).

	10-fold cross validation			Leave-one-out cross validation		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
BRCA	0.977	0.977	0.992	0.977	0.961	0.995
DBLC	0.936	1	0.981	0.963	1	0.981
PAAD	0.941	0.983	0.994	0.912	0.978	0.98
PARD	0.955	0.986	0.988	0.925	0.986	0.988
Overall	0.963	0.984	0.99	0.949	0.976	0.989

The sensitivity, specificity, and AUC of the proposed ensemble model for each class on dataset D_1. The last row, Overall, indicates the average value of each measure. The 10-fold cross validation and leave-one-out cross validation were applied to estimate the evaluation measures.

Table 2b
Classification results on D_1 (SVM as the base classifier).

	10-fold cross validation			Leave-one-out cross validation		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
BRCA	0.977	1	0.993	0.977	0.992	0.988
DBLC	1	1	1	1	1	1
PAAD	1	0.989	0.994	0.971	0.994	0.996
PARD	1	1	1	1	0.993	1
Overall	0.991	0.998	0.996	0.962	0.994	0.994

The sensitivity, specificity, and AUC of the proposed ensemble model for each class on dataset D_1. The last row, Overall, indicates the average value of each measure. The 10-fold cross validation and leave-one-out cross validation were applied to estimate the evaluation measures.

3.4. Performance evaluation

Tables 2a and 2b show the classification results of the proposed method on D_1 in terms of sensitivity, specificity, and AUC. Tables 2c and 2d indicate the results on D_2. Table 2a, shows that the average AUC of the proposed method was 0.99 and 0.989 with the 10-fold cross validation and leave-one-out cross validation, respectively, which clearly indicates that the proposed method with C4.5 as a base classifier provides good prediction on D_1. Similar results with SVM as a base classifier are shown in Table 2b. Tables 2c and 2d shows that the AUC was approximately 0.98 and 0.97, respectively, which implies that the proposed method also works well on D_2. However, the performance on D_2 is slightly lower than that on D_1 because D_2 has fewer samples than D_1. From Table 2, we can easily conclude that the proposed algorithm performs well on both bead-based expression data and miRNA-seq expression data.

We also compared the classification accuracy of our method with C4.5 and SVM as the base classifier, which we denoted as OurC4.5 and OurSVM, respectively, with other commonly used ensemble methods:

Table 2c
Classification results on D_2 (C4.5 as the base classifier).

	10-fold cross validation			Leave-one-out cross validation		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
COLON	0.8	1	0.984	0.8	0.984	0.976
PAN	1	0.969	0.983	1	0.969	0.977
UT	1	1	1	0.9	1	1
T_BALL	0.846	1	0.971	0.846	1	0.978
T_TALL	0.918	0.927	0.982	1	0.927	0.973
Overall	0.918	0.978	0.981	0.904	0.976	0.979

The sensitivity, specificity, and *AUC* of the proposed ensemble model for each class on dataset D_2. The last row, Overall, indicates the average value of each measure. The 10-fold cross validation and leave-one-out cross validation were applied to estimate the evaluation measures.

Table 2d
Classification results on D_2 (SVM as the base classifier).

	10-fold cross validation			Leave-one-out cross validation		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
COLON	0.8	0.984	0.987	0.8	0.984	0.978
PAN	1	0.969	0.99	1	0.969	0.979
UT	0.9	1	1	0.9	1	1
T_BALL	0.846	1	0.971	0.846	1	0.977
T_TALL	1	0.927	0.976	1	0.927	0.969
Overall	0.904	0.976	0.981	0.904	0.976	0.979

The sensitivity, specificity, and AUC of the proposed ensemble model for each class on dataset D2. The last row, Overall, indicates the average value of each measure. The 10-fold cross validation and leave-one-out cross validation were applied to estimate the evaluation measures.

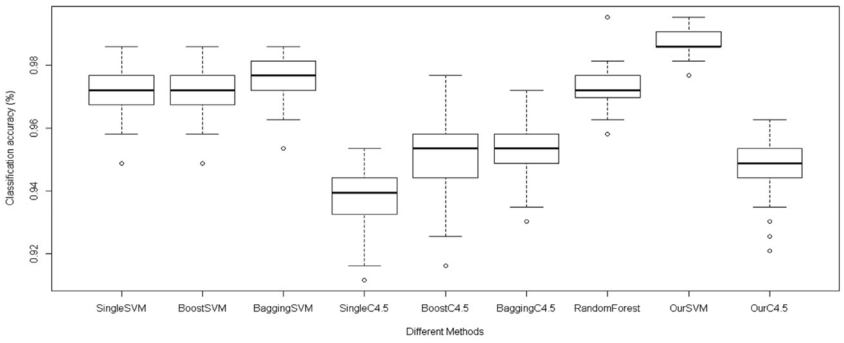
bagging with C4.5 as a base classifier (BaggingC4.5), boosting with C4.5 as a base classifier (BoostC4.5), bagging with SVM as a base classifier (BaggingSVM), boosting with SVM as a base classifier (BoostSVM), and random forests. To achieve impartial results, the same number of classifiers (20) was selected in our experiments for all ensemble methods. Fig. 2a and b show the boxplots of classification accuracies for nine different methods on D_1 and D_2, respectively. Clearly, the proposed method with SVM as a base classifier shows the best accuracy on D_1, which is 0.986 (standard deviation 0.0036), whereas the others are 0.973, 0.973, 0.975, 0.938, 0.951, 0.954, and 0.973 for SingleSVM, BoostSVM, BaggingSVM, SingleC4.5, BoostC4.5, BaggingC4.5, and random forests, respectively. On dataset D_2, the proposed method with C4.5 as a base classifier resulted in the best prediction accuracy, 0.9178 (standard deviation 0.0098), whereas the other methods are 0.8764, 0.8764, 0.8556, 0.8384, 0.8334, 0.8556, and 0.8066. Interestingly, the accuracy of the proposed method is greater when SVM is used as a base classifier on D_1, whereas our method with C4.5 worked better on D_2. The reason is that D_1 has a much larger dimensionality than D_2, and it is well known that SVM can generally achieve higher performance than C4.5 on high-dimensional data.

To examine the statistical significance of the differences among the classifiers, we conducted a paired-sample t-test for BaggingSVM and OurSVM on D_1 and BoostSVM and OurC4.5 on D_2. We selected BaggingSVM and BoostSVM because BaggingSVM had the best performance among the existing methods on D_1, as did BoostSVM on D_2. As a result, the hypothesis that the mean accuracy of the proposed method is equal to the mean accuracy of BaggingSVM on D_1, was soundly rejected (t=−10.186, p-value=0.000). The hypothesis that the mean accuracy of proposed method is equal to the mean accuracy of BoostSVM on D_2 was also soundly rejected (t=−15.9835, p-value=2.2e-16) with a 5% significance level. Consequently, the proposed method significantly outperforms other ensemble methods.

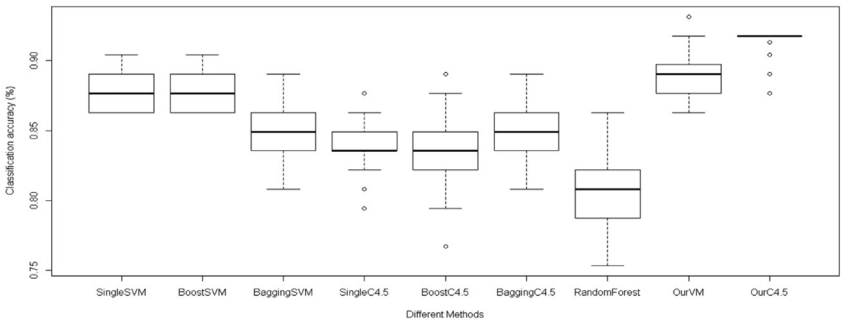
In addition to the classification accuracy, the number of base classifiers is associated with ensemble classifiers, which should also be considered [28]. Fig. 3a and b show the relationships between the number of classifiers and the prediction performance on D_1 and D_2. Fig. 3a and b clearly show that the classification accuracy did not significantly change after a certain number of classifiers. Therefore, it is not a good idea to increase the number of classifiers to as many as possible.

4. Discussion

Ensemble machine learning has been an active research area in the machine learning and bioinformatics community. Many studies [29–31] have shown that ensemble learning is suitable for bioinformatic applications such as gene expression, mass spectrometry-based proteomics data, and gene-gene interaction identification from genome-wide association studies [32]. A single classifier appears unable to work

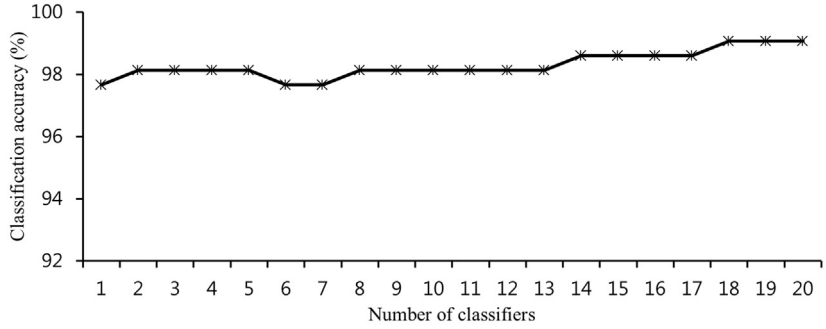


a. Boxplots of the performances of SingleSVM, BoostSVM, BaggingSVM, SingC4.5, BoostC4.5, BaggingC4.5, RandomForest, OurSVM, and OurC4.5 on D_1. Fifty runs of 10-fold cross validation were performed on each classifier.

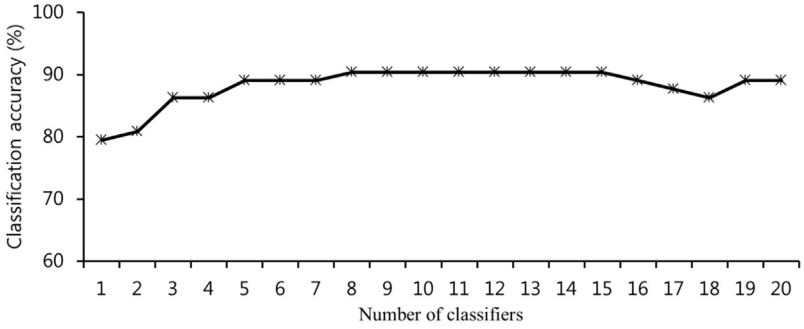


b. Boxplots of the performances of SingleSVM, BoostSVM, BaggingSVM, SingC4.5, BoostC4.5, BaggingC4.5, RandomForest, OurSVM, and OurC4.5 on D2. Fifty runs of 10-fold cross validation were performed on each classifier.

Fig. 2. a. Boxplots of the performances of SingleSVM, BoostSVM, BaggingSVM, SingC4.5, BoostC4.5, BaggingC4.5, RandomForest, OurSVM, and OurC4.5 on D_1. Fifty runs of 10-fold cross validation were performed on each classifier. **b.** Boxplots of the performances of SingleSVM, BoostSVM, BaggingSVM, SingC4.5, BoostC4.5, BaggingC4.5, RandomForest, OurSVM, and OurC4.5 on D2. Fifty runs of 10-fold cross validation were performed on each classifier..



a. Relationship between the prediction performance and the number of base classifiers on D_1



b. Relationship between the prediction performance and the number of base classifiers on D_2.

Fig. 3. a. Relationship between the prediction performance and the number of base classifiers on D_1. b. Relationship between the prediction performance and the number of base classifiers on D_2.

well because of incomplete and noisy biological data. In particular, it is difficult to capture the true classification boundary using a single learner on datasets with very low sample/dimension ratios. Thus, we propose an efficient ensemble method to classify cancers on miRNA expression data. The experiment results demonstrate that the proposed method works well on both bead-based and sequence-based expression data.

Xu et al. [33] proposed a Default ARTMAP algorithm based on a neural network and combined it with a particle swarm optimization feature-selection algorithm to classify various types of cancer. The study reported that the average accuracy of their method was 85% using one of the datasets in this experiment (D_2). Compared with their results, our method significantly outperforms their classifier. Moreover, their method requires a feature selection procedure, whereas our method does not require the additional effort for feature selection because our method selects the subsets of features during the ensemble construction procedure. To the best of our knowledge, there is no study of ensemble learning for cancer classification with miRNA-seq expression data. The emergence of next-generation sequencing technology has prompted various changes in the biological and medical area and produced a large amount of digital miRNA expression data with slightly different characteristics compared with array-based data [34]. Hence, a newly devolved machine learning method should consider the performance on such a dataset. From this viewpoint, our method successfully solves the issues.

The proposed method is based on subsets of features. As a result, the ensemble is not suitable for low-dimensional data because the number of generated subsets is limited with such data. In other words, the algorithm cannot generate a sufficiently high number of base classifiers. However, our method is designed to solve the cancer classification problem with miRNA expression profiles. Unknown miRNAs continue to be discovered by many researchers, and there is no doubt that the dimensionality of the miRNA expression dataset grows daily. Therefore, our method will remain effective when the dimensionality of the dataset consistently grows.

5. Conclusion

In this article, we proposed a feature subset-based ensemble method to classify multiple cancers using miRNA expression data. To generate multiple subsets, we considered the feature relevance and redundancy. Hence, we can generate independent feature subsets and make each subset contain its own information for the classification task. We used a C4.5 decision tree algorithm and an SVM as the base classifiers, with the average probability as the combination method. In the experiments, we used freely accessible expression datasets and tested our algorithm with both 10-fold and leave-one-out cross validation. We also tested the performance by varying the number of base classifiers. The results show that our model has higher prediction accuracy than popular ensemble methods.

Conflicts of interest

None Declared

Acknowledgements

This study was carried out with the support of R & D Program for Forestry Technology (Project No. 2016S1100021) provided by Korea Forest Service and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-H8501-16-1013) supervised by the IITP (Institute for Information & communication Technology Promotion).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiomed.2016.11.008.

References

[1] Y.J. Kim, H.Y. Yoon, J.S. Kim, et al., HOXA9, ISL1 and ALDH1A3 methylation patterns as prognostic markers for nonmuscle invasive bladder cancer: array-based DNA methylation and expression profiling, Int. J. Cancer 133 (2013) 1135–1143.
[2] S. Kar, K.D. Sharma, M. Maitra, Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique, Expert Syst. Appl. 42 (2015) 612–627.
[3] T. Latkowski, S. Osowski, Data mining for feature selection in gene expression autism data, Expert Syst. Appl. 42 (2015) 84–872.
[4] X. Wang, Robust two-gene classifiers for cancer prediction, Genomics 99 (2011) 90–95.
[5] T. Abeel, T. Helleputte, P.Y. Van., et al., Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (2010) 392–398.
[6] S. Sassen, E.A. Miska, C. Caldas, MicroRNA–implications for cancer, Virchows Arch. 452 (2008) 1–10.
[7] S. Mi, J. Lu, M. Sun, et al., MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia, PNAS 104 (2007) 19971–19976.
[8] G.A. Calin, C.M. Croce, MicroRNA signatures in human cancers, Nat. Rev. Cancer 6 (2006) 857–866.
[9] J. Lu, G. Getz, E.A. Miska, et al., MicroRNA expression profiles classify human cancers, Nat. Rev. Cancer 6 (2005) 857–866.
[10] E. Fridman, Z. Dotan, I. Barshack, et al., Accurate molecular classification of renal tumors using microRNA expression, J. Mol. Diagn. 12 (2010) 687–696.
[11] E.J. Nam, H. Yoon, S. Kim, et al., MicroRNA expression profiles in serous ovarian carcinoma, Clin. Cancer Res. 14 (2008) 2690–2695.
[12] G. Zenobi, P. Cunningham, Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error, Mach. Learn.: ECML 2167 (2001) 576–587.
[13] A. Rahman, B. Verma, Ensemble classifier generation using non-uniform layered clustering and genetic algorithm, Knowl. Based Syst. 43 (2013) 30–42.
[14] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (2010) 1–39.
[15] Z. Zeng., H. Zhang, R. Zhang., et al., A novel feature selection method considering feature interaction, Pattern Recognit. 48 (2015) 2656–2666.
[16] Y. Piao, M. Piao, K. Park, et al., An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data, Bioinformatics 28 (2012) 3306–3315.
[17] S.S. Kannan, N. Ramaraj, A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm, Knowl. Based Syst. 23 (2010) 580–585.
[18] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: Proceedings of the Twentieth International Conference on Machine Learning, Washington DC, 856, 2003.
[19] J.R. Quinlan, C4. 5: Programs forMachine Learning, 1993.
[20] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[21] L.I. Kuncheva, A theoretical study on six classifier fusion strategies, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2004) 281–286.
[22] P. Clark, R. Boswell, Rule induction with CN2: Some recent improvements, Mach. Learn. (1991) 151–163.
[23] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[24] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[25] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, Int. Conf. Mach. Learn. (1996) 148–156.
[26] M. Hall, E. Frank, G. Holmes, et al., The WEKA data mining software: an update, SIGKDD Explor. 11 (2009).
[27] S. Ma, J. Huang, Regularized ROC method for disease classification and biomarker selection with microarray data, Bioinformatics 21 (2005) 4356–4362.
[28] H. Liu, L. Liu, H. Zhang, Ensemble gene selection for cancer classification, Pattern Recognit. 43 (2010) 2763–2772.
[29] M. Hilario, A. Kalousis, Approaches to dimensionality reduction in proteomic biomarker studies, Brief. Bioinform. 9 (2008) 102–118.
[30] Y.A. Meng, Y. Yu, L.A. Cupples, L.A. Farrer, K.L. Lunetta, Performance of random forest when SNPs are in linkage disequilibrium, BCM Bioinform. 10 (2009) 78.
[31] M.E.A. Bashir, K.S. Ryu, U. Yun, K.H. Ryu, Pro-detection of atrial fibrillation using mixture of experts, IEICE Trans. Inf. Syst. E95 (2012) 2982–2990 D.
[32] P. Yang, J. Ho, A.Y. Zomaya, B.B. Zhou, A genetic ensemble approach for gene-gene interaction identification, BMC Bioinform. 11 (2010) 254.
[33] R. Xu, J. Xu, D.C. Wunsch, MicroRNA expression profile based cancer classification using Default ARTMAP, Neural Netw. 22 (2009) 774–780.
[34] P. Li, Y. Piao, H.S. Shon, K.H. Ryu, Comparing the normalization methods for the differential analysis of illumina high-throughput RNA-seq data, BMC Bioinform. 16 (2015) 347.