

CLASSIFIER FUSION FOR POORLY-DIFFERENTIATED TUMOR CLASSIFICATION USING BOTH MESSENGER RNA AND MICRORNA EXPRESSION PROFILES

Yuhang Wang* and Margaret H. Dunham

*Department of Computer Science and Engineering, Southern Methodist University,
Dallas, TX 75205, USA*

**Email: yuhangw@engr.smu.edu*

James A. Waddle

*Department of Biology, Southern Methodist University,
Dallas, TX 75205, USA*

Monnie McGee

*Department of Statistical Science, Southern Methodist University,
Dallas, TX 75205, USA*

MicroRNAs (miRNAs) are an important class of small non-coding RNAs that regulate diverse biological processes. MiRNAs are thought to regulate gene expression by degrading or repressing target messenger RNAs (mRNAs) at the post-transcriptional level. Recent studies suggest that miRNAs are implicated in human cancers. In a recent paper, Lu *et al.* showed that the expression profile of 217 mammalian miRNAs could be used to successfully classify poorly differentiated tumor samples at the accuracy of 70.6%, whereas the same classifier using mRNA profiles resulted in a low accuracy of 5.9%. Because miRNAs regulate gene expression at the post-transcriptional level, we hypothesize that miRNA expression profiles can provide information that is complementary to mRNA expression profiles. Therefore, a data fusion approach could lead to improved classification accuracy. As a proof of concept, we re-analyzed the data in the paper by Lu *et al.* using a classifier fusion approach that utilizes both mRNA and miRNA expression data. We built a meta-classifier from two bagged k-nearest-neighbor classifiers. Experimental results showed that our meta-classifier was able to classify the same set of poorly differentiated tumor samples at an improved accuracy of 76.5%, when trained only with the expression profiles of more-differentiated tumor samples.

1. INTRODUCTION

MicroRNAs (miRNAs) are an important class of small non-coding RNAs that regulate diverse biological processes³. MiRNAs are thought to regulate gene expression by degrading or repressing target messenger RNAs (mRNAs) at the post-transcriptional level¹. They have been shown to control cell growth, differentiation and apoptosis^{3, 1}. Consequently, recent studies suggest that miRNAs are implicated in human tumorigenesis^{5, 8} and differentially expressed in cancers^{17, 20}. Most interestingly, Lu *et al.*¹⁷ recently showed that the expression profile of 217 mammalian miRNAs could be used to successfully classify poorly differentiated tumor samples at the accuracy of 70.6%, whereas the same classifier using messenger RNA profiles (microarray gene expression data) resulted in a low accuracy of 5.9%.

On the other hand, significant work has been done in cancer classification based on microarray gene expression data^{9, 2, 22}. Because miRNAs regulate gene expression at the post-transcriptional level, we hypothesize that miRNA expression profiles can provide information that is complementary to mRNA expression profiles. Therefore, a data fusion approach could lead to improved classification accuracy. To investigate this issue, we propose a meta-classifier that uses both miRNA and mRNA expression profiles to classify poorly differentiated tumor samples. As a proof of concept, we then re-analyze the data set used in the paper by Lu *et al.*¹⁷ using our approach.

The remainder of the paper is organized as follows. We give a brief review of the data fusion methods in the literature in the next section. In Section 3, we describe our proposed classifier fusion approach.

*Corresponding author.

Section 4 presents the experiment results on the data set used in the paper by Lu *et al.*¹⁷. Section 5 concludes the paper.

2. RELATED WORK ON DATA FUSION

Data fusion is a well-studied subject in machine learning. By means of data fusion, different sources of information are combined to improve the performances of a system. Using an appropriate fusion scheme, one may expect improved classification accuracy due to the use of complementary information. Data fusion processes are often categorized in three levels, depending on the processing stage at which fusion takes place:

- (1) Low-level fusion¹¹ combines several sources of raw data to produce new raw data that is expected to be more informative and synthetic than the inputs. This kind of fusion is not feasible for integrating mRNA and microRNA expression data.
- (2) Intermediate-level fusion or feature-level fusion^{24, 10}. Here various features extracted from several sources of raw data are combined into a composite feature that may then be used by further processing stages. In the context of cancer classification using both mRNA and microRNA expression data, one possible feature-level fusion approach can be a linear concatenation of the mRNA feature vectors (mRNA expression profiles for tissue samples) and the miRNA feature vectors (miRNA expression profiles for tissue samples). However, we tested this approach on data in Lu *et al.*¹⁷ and found that it actually lead to poorer classification performance (data not shown) than what was reported by Lu *et al.*¹⁷ using miRNA expression profiles only.
- (3) High-level fusion^{7, 12} (or decision fusion, classifier fusion) in which each source of input yields a decision and the decisions are fused. The classifiers may return a “soft” class label and not a “hard” decision. To distinguish both cases, one speaks of hard and soft fusion. Methods of decision fusion¹⁶ include voting methods, statistical methods, fuzzy logic based methods, etc. The

classifier fusion approach we propose in this paper falls under this category.

In our data fusion problem, we only have two data sources (miRNA and mRNA expression profiles). Therefore, simple classifier fusion schemes such as majority voting cannot be used here. Our proposed meta-classifier utilizes two bagged fuzzy k-nearest-neighbor classifiers and picks the output of the more confident one as the final classification output. Thus our approach is well-suited for integrating two data sources. We described the proposed approach in detail in the next section.

3. CLASSIFIER FUSION FOR mRNA AND miRNA EXPRESSION DATA INTEGRATION

An overview of our proposed classifier fusion system is illustrated in Figure 1. In this approach, we first apply the Relief-F feature selection algorithm¹⁵ to the training data to select top-ranked informative genes from mRNA expression data and miRNA expression data separately. Then, using only the selected genes as features, for each of the two types of expression data, a fuzzy k-nearest-neighbor (k-NN) classifier augmented with bagging is trained on the training data and applied to the test data. For each tissue sample in the test data set, the classification outputs (soft class labels) from the two classifiers are compared, and the output from the classifier with better confidence is then chosen as the final class assignment.

3.1. Gene Selection using Relief-F

Relief-F¹⁵ is one of the most widely used feature selection algorithm. The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. Specifically, it tries to find a good estimate of the following probability to assign as the weight for each feature f .

$$w_f = P(\text{different value of } f | \text{different class}) - P(\text{different value of } f | \text{same class})$$

This approach has shown good performance

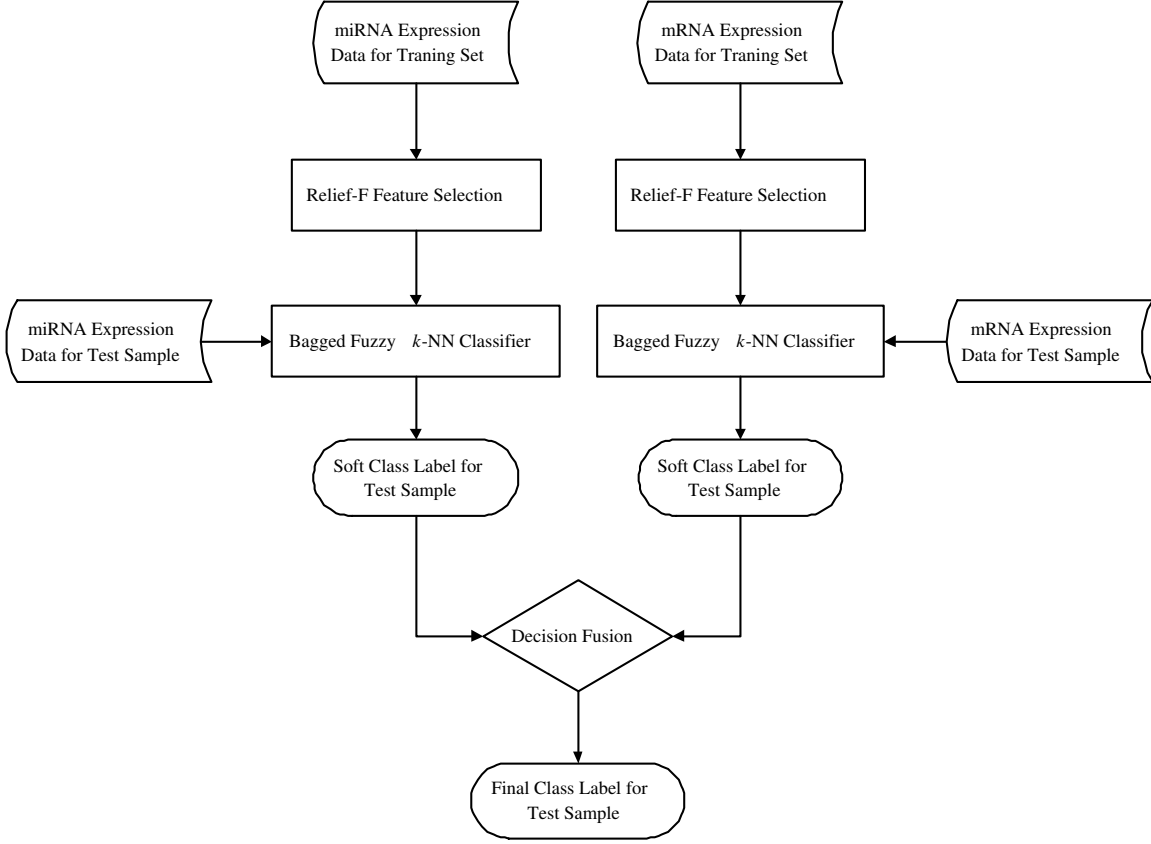


Fig. 1. Classifier fusion system for cancer classification.

in various domains¹⁹, including informative gene selection²¹.

3.2. Fuzzy k -NN Classifiers

Here we chose to use the fuzzy k -NN classifier because it supports non-linear decision boundaries and is naturally applicable to multi-class classification problems. The k -NN classifier⁶ is a well-known non-parametric classifier. Fuzzy k -NN extends k -NN by replacing the crisp class labels with *soft labels*, $l(v_i) \in [0, 1]^c$. Different ways of combining the soft labels of the k nearest neighbors to compute the soft output label $l(x)$ for $x \in X$ have been proposed^{13, 14, 6}. In this work, we use the scheme by Keller *et al.*¹⁴. To find the i th component of the soft class label $l_i(x)$ for x , the method by Keller *et al.*¹⁴ takes distances into consideration:

$$l_i(x) = \frac{\sum_{j=1}^k l_i(z_j)(d_j)^{-\frac{2}{m-1}}}{\sum_{j=1}^k (d_j)^{-\frac{2}{m-1}}} \quad (1)$$

where

m is a “fuzzification” parameter;

d_j is the distance between x and its j th nearest neighbor z_j .

In our case, however, the class labels of the training data are actually crisp labels. If a training sample z belongs to class i , the i th component of the soft class label $l_i(z)$ for z is 1, and all other components are 0.

3.3. Bagging

Bagging classifiers⁴ is a method for generating multiple versions of a classifier and using them to obtain an aggregated classifier. It is also used to address the inherent instability of results when applying classifiers to relatively small data sets. Here, multiple versions of the fuzzy k -NNa classifier are obtained by repeatedly sub-sample (with replacement) from the training data. The outputs (soft class labels) from those classifiers are then simply averaged.

Table 1. Comparison of classification accuracies and P values.

	Classification accuracy	P value
Classifier fusion	76.5%	1.4×10^{-12}
Bagged fuzzy k -NN classifier using miRNA expression data	70.6%	4.8×10^{-11}
Bagged fuzzy k -NN classifier using mRNA expression data	47.1%	5.2×10^{-6}
Original approach using miRNA expression data	70.6%	4.8×10^{-11}
Original approach using mRNA expression data	5.9%	0.47

3.4. Decision fusion

Two bagged fuzzy k -NN classifiers are trained with mRNA and miRNA expression data separately. For each tissue sample in the test data set, the decision of the two classifiers are aggregated by choosing the decision of the classifier with higher confidence. Note that we do not average the outputs from the two classifiers. One of the two classifiers is actually chosen based on its confidence. Here, we use the highest value in the soft class label as the degree of confidence of the classifier about the class assignment of a test sample.

4. RESULTS

4.1. Data

We downloaded the processed miRNA expression data and the corresponding mRNA expression data (published in Ref. 18) from <http://www.broad.mit.edu/cancer/pub/miGCM>.

The data has two parts: a training set of 68 more differentiated tumors, representing 11 tumour types; and a test set of 17 poorly differentiated test samples.

Each sample was profiled in the space of 217 mammalian miRNAs and $\sim 16,000$ mRNAs.

4.2. Experiment Setup

For the mRNA expression data, top 40 genes were selected using Relief-F algorithm. For the miRNA expression data, top 40 miRNA genes were selected using Relief-F. In both cases, 5 nearest neighbors were used in Relief-F for estimating feature weights. For the bagging algorithm, the number of bagging iterations was 10 and the size of each bag was the same as the size of the training data. For the fuzzy k -NN classifiers, we used the Euclidean distance as the distance metric in the experiments, and the best k between 1 and 5 was found by performing leave-

one-out cross-validation on the training data.

The system was implemented using Perl and Weka 3.4.7²³.

4.3. Empirical Results

Table 1 shows the classification accuracies and P values of our approach and the original approaches used by Luet *et al.*¹⁷. We can observe that our classifier fusion approach was able to classify the same set of poorly differentiated tumor samples at an improved accuracy of 76.5%, as compared to the accuracy of 70.6% obtained using the original approach (a probabilistic neural network algorithm) described in the paper by Lu *et al.*¹⁷. The bagged fuzzy k -NN classifier using only miRNA expression data achieved the same classification accuracy of 70.6% as the original approach. Interestingly, the bagged fuzzy k -NN classifier using only mRNA expression data achieved a classification accuracy of 47.1% ($P = 5.2 \times 10^{-6}$), far exceeding the classification accuracy of 5.9% ($P = 0.47$) obtained by the original approach on the same data.

5. CONCLUSIONS

In this paper, a classifier fusion approach is proposed for poorly-differentiated tumor classification using both mRNA and miRNA expression profiles. Preliminary results on a public data set showed improved classification accuracy compared to that obtained using only mRNA or miRNA expression profile. We will test the proposed classifier fusion approach on more data sets when they become available.

References

1. AMBROS, V. The functions of animal microRNAs. *Nature* 431, 7006 (2004), 350–355.
2. ARMSTRONG, S. A., STAUNTON, J. E., SILVERMAN, L. B., PIETERS, R., DEN BOER, M. L., MINDEN,

- M. D., SALLAN, S. E., LANDER, E. S., GOLUB, T. R., AND KORSMEYER, S. J. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 30, 1 (2002), 41–47.
3. BARTEL, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 2 (2004), 281–297.
 4. BREIMAN, L. Bagging predictors. *Machine Learning* 24, 2 (1996), 123–140.
 5. CROCE, C. M., AND CALIN, G. A. miRNAs, cancer, and stem cell division. *Cell* 122, 1 (2005), 6–7.
 6. DASARATHY, B. *Nearest Neighbor Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, 1991.
 7. DASARATHY, B. *Decision Fusion*. Computer Society Press, Los Alamitos, California, 1994.
 8. ESQUELA-KERSCHER, A., AND SLACK, F. J. Oncomirs – microRNAs with a role in cancer. *Nat Rev Cancer* 6, 4 (2006), 259–269.
 9. GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLIER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D., AND LANDER, E. S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 5439 (1999), 531–537.
 10. GUNATILAKA, A., AND BAERTLEIN, B. Feature-level and decision-level fusion of noncoincidentally sampled sensors for land mine detection. *PAMI* 23, 6 (June 2001), 577–589.
 11. HALL, D., AND LINAS, J. *Handbook of Multisensor Data Fusion*. CRC Press, 2001.
 12. HO, T., HULL, J., AND SRIHARI, S. Decision combination in multiple classifier systems. *PAMI* 16, 1 (January 1994), 66–75.
 13. JÓZWIK, A. A learning scheme for a fuzzy k-nn rule. *Pattern Recognition Letters* 1 (1983), 287–289.
 14. KELLER, J. M., GRAY, M. R., AND GIVENS, J. A. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man and Cybernetics* 15, 4 (1985), 580–585.
 15. KONONENKO, I. Estimating attributes: analysis and extensions of relief. In *Proceedings of the European conference on machine learning on Machine Learning* (1994), Springer-Verlag New York, Inc., pp. 171–182.
 16. KUNCHEVA, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.
 17. LU, J., GETZ, G., MISKA, E. A., ALVAREZ-SAAVEDRA, E., LAMB, J., PECK, D., SWEET-CORDERO, A., EBERT, B. L., MAK, R. H., FERRANDO, A. A., DOWNING, J. R., JACKS, T., HORVITZ, H. R., AND GOLUB, T. R. MicroRNA expression profiles classify human cancers. *Nature* 435, 7043 (2005), 834–838.
 18. RAMASWAMY, S., TAMAYO, P., RIFKIN, R., MUKHERJEE, S., YEANG, C. H., ANGELO, M., LADD, C., REICH, M., LATULIPPE, E., MESIROV, J. P., POGGIO, T., GERALD, W., LODA, M., LANDER, E. S., AND GOLUB, T. R. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 98, 26 (2001), 15149–54.
 19. ROBNIK-SIKONJA, M., AND KONONENKO, I. Theoretical and empirical analysis of relief and rrelief. *Mach. Learn.* 53, 1-2 (2003), 23–69.
 20. VOLINIA, S., CALIN, G. A., LIU, C. G., AMBS, S., CIMMINO, A., PETROCCA, F., VISONE, R., IORIO, M., ROLDO, C., FERRACIN, M., PRUEITT, R. L., YANAIHARA, N., LANZA, G., SCARPA, A., VECCHIONE, A., NEGRINI, M., HARRIS, C. C., AND CROCE, C. M. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A* 103, 7 (2006), 2257–2261.
 21. WANG, Y., AND MAKEDON, F. Application of relief-f feature filtering algorithm to selecting informative genes for cancer classification using microarray data. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference* (Stanford, California, 2004), pp. 477–478.
 22. WANG, Y., MAKEDON, F. S., FORD, J. C., AND PEARLMAN, J. HyKGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics* 21, 8 (2005), 1530–1537.
 23. WITTEN, I. H., AND FRANK, E. *Data mining : practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, Calif., 1999.
 24. YANG, J., YANG, J., ZHANG, D., AND LU, J. Feature fusion: parallel strategy vs. serial strategy. *Pattern Recognition* 36, 6 (June 2003), 1369–1381.