Integrating MicroRNA and mRNA Expression Data for Cancer Classification

Hasan Oğul and Onur Altındağ
Department of Computer Engineering, Baskent University, Ankara, Turkey

Keywords: Tumor Classification, Gene Expression, Data Integration, Feature Selection, MicroRNA Regulation.

Abstract: Classifying cancer samples from gene expression data is one of the central problems in current systems biomedicine. The problem is challenging due to the small number of samples in comparison to the number of genes (mRNAs) in a typical microarray experiment. Recent reports suggest that feature selection may help to manage the problem. Furthermore, microRNA expression profiles have shown to provide valuable knowledge in detecting cancer signatures. In this study, we present the results of a comprehensive study to assess the effect of feature selection and microRNA-mRNA data integration in cancer type prediction from microarray expression data. We prove that this integration can significantly improve prediction accuracy with a proper feature selection strategy.

1 INTRODUCTION

One of the most challenging issues in current data analysis science is so-called "small *n*, large *p*" paradigm, where *n* is the number of samples and *p* is the number of features in present data to be analyzed. To address this issue, three main strategies come into prominence in the literature; feature selection, data integration and probabilistic modelling. Feature selection is proper removal of a set of features which have probably no or less putative effect on inference. In this approach, the *p* is simply reduced by selecting a pretty small subset of all features (Saeys et al., 2007). Data integration (or fusion) is defined as exploiting multiple sources of data which may help to improve final decision made. This integration may appear in several ways such as using additional measurements (samples) or considering other factors which may have complementary effects on the original features (Huopaniemi et al., 2010); (Ogul and Akkaya, 2011). Probabilistic modelling approach is built over a Bayesian assumption for inference. The methods in this category encode and manipulate probability distributions to model the uncertainty over high-dimensional spaces (West, 2003); (Klami and Kaski, 2008).

In current systems biomedicine, an important problem is to diagnose the type of a tumor from a given diseased tissue sample. A tissue sample is usually accompanied with a set of mRNA expression profiles obtained from microarray experiments. Since each cancer type distinctively alters the regulatory behaviours of some related genes, these profile sets have a strong potential to identify tumor types. In this set, each tissue sample is represented by a fixed number of expression values which correspond to the activities of all known genes in the genome. The problem then turns out to be a pattern classification task where a vector of gene expression values is required to be assigned to one of the known classes of cancer. Since the number of these samples are often too less in comparison with the number of all genes, "small *n* large *p*" problem frustratingly appears here. In the last decade, various machine learning techniques have been used to improve the prediction accuracy of cancer classifiers (Ramaswamy et al., 2001); (Su et al., 2001); (Su et al., 2003); (Peng et al., 2003); (Lin et al., 2006); (Xu et al., 2007); (Peng et al., 2009); (Liu and Xu, 2009). While there have been these developments in machine learning site, we have witnessed a drastic shift in understanding of gene regulation in biosciences. It has been proven that some tiny molecules, called microRNAs, have additional complementary or pivotal effects on gene regulatory networks. It is now evidently known that they take important roles in regulation of thousands of gene in post-transcriptional level (Bartel, 2004). Some

Oğul H. and Altındağ O. (2013).
Integrating MicroRNA and mRNA Expression Data for Cancer Classification.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 503-507
DOI: 10.5220/0004334405030507

recent studies have shown that the knowledge of microRNA expression changes alone is very promising in classifying cancer types (Lu et al., 2005); (Xua et al., 2009); (Chan et al., 2011).

In this study, we attempt to integrate the knowledge of microRNA regulation and the benefits of data fusion and feature selection strategies to overcome "small *n* large *p*" problem in the task of cancer classification from expression data. To this end, we deploy five well-known machine learning algorithms with five distinct feature selection criteria for multi-category cancer classification. According to the experimental results on a common benchmark set, the integration of mRNA and microRNA expression data can remarkably improve the prediction accuracy of cancer classifiers provided that a proper feature selection strategy is employed. To the authors' knowledge, the best accuracy ever is reported on a benchmark dataset.

2 MATERIALS AND METHODS

2.1 Classification

The main problem is to assign an unknown tissue sample to one of the given cancer categories including normal type. Several machine learning algorithms exist for multi-category classification in the literature. Considering their common use in systems biology and reported success in other domains, we choose five among these algorithms and evaluate their classification performance on selected datasets: C4.5 Decision Tree (DT), Artifical Neural Networks (ANN), Support Vector Machines (SVM), Naïve Bayes multinomial classifier (NBM), and K-Nearest Neighbors (KNN). A brief summary and comparison of these methods can be found in Caruana and Niculescu-Mizil (2006). For their detailed descriptions, the author is referred to Bishop (2006).

2.2 Feature Selection

Feature selection is the task of creating a reduced and possibly more informative subset of all features over whole samples of given data. It is proven to be a critical need for mRNA data to get accurate and trustworthy tumor classification results (Guyon et al., 2002); (Cai et al., 2007). We assessed several methods for feature selection in terms of their previous performances on similar problems and chosen five among these algorithms: SVM attribute selection, Information Gain based attribute selection, Gain Ratio based attribute selection, chi-squared test-based feature selection and CFS subset attribute selection (Saeys et al., 2007); (Guyon et al., 2002); (Hall, 1998). All feature selection methods were run using their default parameters in Weka, the machine learning tool that we used in our experiments (Hall et al., 2009).

2.3 Data Sets

We use mRNA and microRNA expression profiles from paired samples of normal and diseased tissues with different cancer types: colon, pancreas, kidney, bladder, prostate, ovary, uterus, lung, meso, mela and breast. Individual and integrated mRNA and microRNA datasets are organized as follows:

mRNA expression profiles dataset: This is a subset of GCM (Global Cancer Map) mRNA dataset provided by Ramaswamy et al. (2001). It contains 89 samples with the expression profiles of 16,063 genes from 11 classes of tumors and some normal samples for each tissue. All the normal tissue samples were grouped in a single "normal" class.

miRNA expression profiles dataset: Lu et al. (2005) used a bead-based flow cytometric miRNA expression profiling method to present a systematic expression analysis of 217 mammalian miRNAs from the same samples as Ramaswamy et al. used. We use a subset of this miRNA dataset containing the same 89 samples as *mRNA expression profiles dataset* with 217 miRNAs.

miRNA & mRNA expression profiles dataset: This is the combination set of *miRNA expression profiles dataset* and *mRNA expression profiles dataset* with 16,280 features in total.

2.4 Experimental Setup

We performed 90 experiments by compiling five classifiers on three datasets with their original and reduced versions that are result of the five feature selection methods mentioned above. In the experiments, we used LOOCV (Leave-one-out cross validation) technique to test the accuracies of the classifiers on selected datasets. It is well known that this validation technique gives the most realistic accuracy results for "small *n*, large *p*" experiments. These kinds of experiments can result with over fitting if a proper and sufficient validation on training is not performed. The accuracy is simply defined as the percentage of correctly classified samples.

Peng et al. (2009) performed a similar multi-

cancer classification study with the same datasets over the same experimental setup. By using LOOCV, they reported realistic and reproducible comparisons with previous studies. According to their results, they mainly argued that microRNA information alone is not sufficient to classify many types of cancer but the mRNA information alone is, whereas Lu et al. (2005) argued the opposite. Peng et al. supports the need of effective feature selection/reduction on the data used and successfully demonstrate the classifier performance with mRNA profile is superior to that with microRNA profile. In our study we suggest that neither the claim of Lu et al. (2005) nor Peng et al. (2009) is wrong but inadequate. Both microRNA and mRNA information is very valuable for tumor classification. It is also scientifically proven that they are related to each other. Based upon this fact we suggest that, by effective fusion of these data and optimized use of machine learning algorithms and feature selection methods, it is possible to achieve better prediction accuracy for multi-class tumor classification problems.

Since the machine learning classifiers are usually very sensitive to the initial parameter sets defined, we used some greedy optimization algorithms and user assisted methods to find better parameter combinations for the classifiers that work with different parameter options. This parameter selection is applied for KNN, ANN and SVM, while the default parameters set in Weka is compiled for C4.5 and NBM.

We propose a winner-score based system to evaluate the general discriminative ability of microRNA-mRNA data integration. For each pair of classification and selection method, we noted the best accuracy among three datasets (sole miRNA, sole mRNA or their combination). At each run, we incremented the related score for winning data set to get a general winner-score, such that the maximum score would be 30. Overall evaluation of this scoring scheme is expected to show literally if the fusion of these two biological data can lead to better multi-tumor classification results or not, in general.

3 RESULTS

Experimental results are shown in Table 1. We first evaluated the classification performances of miRNA, mRNA and the fusion set without feature selection with five selected algorithms. The results indicate that the classification performances with sole microRNA data are all better than the others with sole mRNA data as it was mentioned by Lu et al. (2005). However, the fusion set even performed slightly better for some classifiers (C4.5 decision trees and SVM).

Next, we performed the same tests with five feature selection methods applied on all three datasets. The results we got support the findings by Peng et al. (2009), i.e. feature-selected mRNA data mostly yields better accuracy than microRNA data. Nevertheless, combined dataset got a winner-score of 26 (out of 30), showing that it outperformed both

Table 1: LOOCV accuracy comparison results of the multi-cancer classification experiments performed.

Feature selection method	Dataset (with number of selected features fed to classifier)	LOOCV accuracy (%) with different classifiers					Winner-scores of single and combined datasets		
		KNN	ANN	DT	NBM	SVM	mRNA	miRNA	miRNA&mRNA
No Attribute Selection	mRNA (16063 features)	60,7	23,6	38,2	55,1	**75,3**	**0/5**	**3/5**	**2/5**
	miRNA (217 features)	68,5	**83,1**	51,7	75,3	77,5			
	miRNA & mRNA (16280 features)	60,7	23,6	52,8	57,3	**77,5**			
SVM Attribute Selection	mRNA (100)	88,8	**95,5**	41,6	85,4	92,1	**0/5**	**0/5**	**5/5**
	miRNA (100)	73,0	**86,5**	46,1	75,3	82,0			
	miRNA & mRNA (100)	92,1	**96,6**	70,8	91,0	93,3			
Information Gain Attribute Selection	mRNA (365)	80,9	**89,9**	53,9	75,3	84,3	**1/5**	**0/5**	**4/5**
	miRNA (76)	73,0	**83,1**	40,4	70,8	82,0			
	miRNA & mRNA (441)	85,4	88,8	67,4	87,6	**88,8**			
Gain Ratio Attribute Selection	mRNA (<365)	80,9	**89,9**	55,1	75,3	84,3	**0/5**	**0/5**	**5/5**
	miRNA (<76)	76,4	**85,4**	40,4	70,8	84,3			
	miRNA & mRNA (<441)	87,6	**92,1**	67,4	87,6	88,8			
Chi-Squared Attribute Selection	mRNA (<365)	80,9	**89,9**	56,2	77,5	84,3	**0/5**	**0/5**	**5/5**
	miRNA (76)	73,0	**83,1**	40,4	70,8	82,0			
	miRNA & mRNA (<=441)	85,4	**89,9**	69,7	87,6	88,8			
CFS Subset Attribute Selection	mRNA (90)	88,8	**93,3**	55,1	84,3	91,0	**0/5**	**0/5**	**5/5**
	miRNA (18)	68,5	74,2	55,1	46,1	71,9			
	miRNA & mRNA (91)	88,8	93,3	67,4	89,9	**93,3**			
TOTAL SCORE							**1/30**	**3/30**	**26/30**

sole microRNA and sole mRNA datasets in general. When we examined the results, we also noticed that every selection method we used tends to chose a mixture of features as best features from both miRNA and mRNA but never from only one.

The best classification accuracy we obtained is 96.6% (ANN with SVM Attribute selection). This result is better than the best LOOCV performance on the same dataset in the literature, which was reported as 95.8% by Peng et al. (2009). LOOCV accuracy comparison of multi-class classification using the GCM datasets is given in Table 2. In addition to performance comparison of the datasets, we compared the performances of the classifiers on these experiments. According to the results ANN performed best followed by SVM. But SVM have a larger optimization capacity and much faster training performance so they can yield better accuracies in our future work.

Table 2: Comparison with other results (LOOCV accuracy of cancer classification on GCM datasets).

Studies	Accuracy (%)
Ramaswamy et al., 2001	78.0
Su et al., 2003	81.3
Peng et al., 2003	85.2
Lin et al., 2006	84.3
Liu and Xu, 2009	91.8
Peng et al., 2009	95.8
This study	**96.6**

4 CONCLUSION AND FUTURE WORK

We evaluate to what extend the integration of microRNA and mRNA expression data can improve the prediction accuracy of multi-category cancer classifiers. Based on the results of a rigorous experimental study, we have shown that with proper feature selection strategies, the integration of microRNA and mRNA data by feature-level fusion can significantly improve the prediction performance and provide better classification accuracy than single use of mRNA and microRNA data.

Later on this study, we will continue to optimize the feature selection and classification methods for better accuracy. We will also be working with different datasets comprising paired microRNA and mRNA expression profiles over diseased samples. We will especially focus on predicting subtypes of vital cancers. Another future direction is to compare the performance of potential knowledge-driven feature selection methods with data-driven methods used here. We aim to come up with an integrated hybrid solution for cancer classification and provide a web server for the use of biomedical researcher working in this domain.

ACKNOWLEDGEMENTS

This study was supported by the Scientific and Technological Research Council of Turkey (TUBİTAK) under the Project 110E160.

REFERENCES

Bartel, D. P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116, 281-297.

Bishop C. M., 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, NJ, USA.

Cai, Z., Goebel, R., Salavatipour, M. R., and Lin, G., 2007. Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. *BMC Bioinformatics*, 8, 206.

Caruana, R., Niculescu-Mizil, A., 2006, An Empirical Comparison of Supervised Learning Algorithms, *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA.

Chan, E., Patel, R., Nallur, S., Ratner, E., Bacchiocchi, A., Hoyt, K., Szpakowski, S., Godshalk, S., Ariyan, S., Sznol, M., Halaban, R., Krauthammer, M., Tuck, D., Slack, F.J., Weidhaas, J.B., 2011. MicroRNA signatures differentiate melanoma subtypes, *Cell Cycle,* 10, 1845-1852.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning,* 46, 389-422.

Hall, M. A., 1998. Correlation-based Feature Subset Selection for Machine Learning. *PhD Thesis*, Hamilton, New Zealand.

Huopaniemi, I., Suvitaival, T., Nikkilä, J., Orešič, M., Kaski, S., 2010. Multivariate multi-way analysis of multi-source data. *Bioinformatics,* 26, i391-i398.

Klami, A., Kaski, S., 2008. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72, 39-46.

Lin, T. C., Liu, R. S., Chen, C. Y., Chao, Y. T., and Chen, S.Y., 2006. Pattern classification in DNA microarray data of multiple tumor types. *Pattern Recognit.*, 39, 2426-2438.

Liu, K. H., and Xu, C. G., 2009. A genetic programming-based approach to the classification of multiclass microarray datasets. *Bioinformatic,s* 25, 331-337.

Liu, K. H., and Xu, C. G., 2009. A genetic programming-

based approach to the classification of multiclass microarray datasets. *Bioinformatics*, 25, 331-337.

Lu, J., Getz, G., Miska, E., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al., 2005. MicroRNA expression profiles classify human cancers. Nature, 435, 83-838.

Ogul, H., Akkaya, M. S., 2011. Data integration in functional analysis of microRNAs. *Current Bioinformatics,* 6, 462-472.

Peng, S., Xu, Q., Ling, X. B., Peng, X., Du, W., and Chen, L., 2003. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Lett,*. 555, 358-362.

Peng, S., Zeng, X., Li. X., Peng, X., Chen, L., 2009. Multi-class cancer classification through gene expression profiles: microRNA versus mRNA. *J. Genet. Genomics*, 36, 409-416.

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., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci.,* 98, 15149-15154.

Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics,* 23, 2507-2517.

Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P.G., Powell, S.M., Moskaluk, C.A., Frierson, H.F., and Hampton, G.M., 2001. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.*, 61, 7388-7393.

Su, Y., Murali, T. M., Pavlovic, V., Schaffer, M., and Kasif, S., 2003. RankGene: Identification of diagnostic genes based on expression data. Bioinformatics 19: 1578-1579.

West M., 2003. Bayesian Factor Regression Models in the Large p, Small n Paradigm. *Bayesian Statistics*, 7, 723-732.

Xu, R., Anagnostopoulos, G. C., and Wunsch, D. C., 2007. Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* 4, 65-77.

Xua, R., Xub, J., Wunsch, D. C., 2009. MicroRNA expression profile based cancer classification using Default ARTMAP, *Neural Networks*, 22, 774-780.

Hall, M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten H.I., 2009, The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, 11,1.