Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Lung Tumor Diagnosis and Subtype Discovery by Gene Expression Profiling

Lu-yong Wang*, Zhuowen Tu

Abstract— The optimal treatment of patients with complex diseases, such as cancers, depends on the accurate diagnosis by using a combination of clinical and histo-pathological data. In many scenarios, it becomes tremendously difficult because of the limitations in clinical presentation and histopathology. To accurate diagnose complex diseases, the molecular classification based on gene or protein expression profiles are indispensable for modern medicine. Moreover, many heterogeneous diseases consist of various potential subtypes in molecular basis and differ remarkably in their response to therapies. It is critical to accurate predict subgroup on disease gene expression profiles. More fundamental knowledge of the molecular basis and classification of disease could aid in the prediction of patient outcome, the informed selection of therapies, and identification of novel molecular targets for therapy.

In this paper, we propose a new disease diagnostic method, Probabilistic Boosting Tree (*PBTree*) method, on gene expression profiles of lung tumors. It enables accurate disease classification and subtype discovery in disease. It automatically constructs a tree in which each node combines a number of weak classifiers into a strong classifier. Also, subtype discovery is naturally embedded in the learning process. Our algorithm achieves excellent diagnostic performance, and meanwhile it is capable of detecting the disease subtype based on gene expression profile.

I. Introduction

THE accurate determination of tumor's site of origin and pathogenesis is important for cancer diagnosis and treatment. In general, pathologists utilize a variety of histological, genetic and immunologic techniques to make site-specific diagnosis. However, current techniques are limited in their ability to distinguish different tumor types. Many specimens are incorrectly classified due to their morphological similarity to other tumor types. Also, a lot of samples remain poorly differentiated and difficult to relate to any known tumor type. Moreover, many heterogeneous diseases consist of various potential subtypes in molecular basis and differ remarkably in their response to therapies. The development of high-throughput biotechnology has made it feasible systematically monitor the biomarkers in complex disease classification and outcome prediction [1] [2]. The emerging technology of gene expression analysis may serve as molecular fingerprint that allow accurate classification of tumor types. The underlying rationale is that the same tumor classes share some expression profile patterns unique to their classes. These molecular fingerprints might reveal new taxonomies.

Previous studies have demonstrated success in discriminating known tumor types from expression profiles by supervised classification techniques, such as linear discriminate analysis (LDA) [3], logistic discriminate method [4], k-nearest neighbor (kNN) [5], Bayesian methods [6] Support Vector Machine (SVM) [7], Boosting [8]. All these methods tried to distinguish known tumor types from expression profiles. However, they can not discover new subtypes. Unsupervised learning approaches, most commonly used in this problem, have the advantage of being impartial to currently accepted classes, but they may reveal a structure that is not biological significant. Most of the recent publications on this issue utilized clustering techniques, such as hierarchical clustering[9], *k*-means[10], minimal spanning tree [11], mixture modeling [12] and self-organization map[13]. These methods do not cluster the subtypes based on discriminative features compared to normal tissues. In this paper, we are motivated to propose a novel joint classification and subtype discovery algorithm in tumor classification based on gene expression profiles of the disease and control. This method is based on a new algorithm, probabilistic boosting tree, which is capable of learning discriminative models for both classification and class discovery. This method is distinguished by its capacity not only to classify diseases from normal controls, but also to detect subclasses within the tumor samples based on their discriminative features. Evaluations are also performed on public available tumor microarray data.

II. Methods

A. Probabilistic Boosting Tree-Based Algorithm

We propose this method based on new learning framework, called "Probabilistic Boosting Tree", which use AdaBoost as the basic unit for learning process. We introduce this method in the logic of AdaBoost and probabilistic boosting tree.

- *AdaBoost*

For self-consistency, we describe the general AdaBoost algorithm first. In general, boosting is a method for improving the accuracy of any given learning algorithm. AdaBoost solved many practical difficulties of earlier boosting methods [14]. It takes $(x1; y1; w1)\ldots\ (x_n; y_n; w_n)$ as input, where each *xi* belongs to some instance space (gene expression profiles in this case), y_i belongs to label set $Y \in \{+1,-1\}$ (disease or control), and w_i is the weights of the samples. AdaBoost calls a given base learning algorithm repeated in t rounds. $D_t(i)$ represents the weight of the distribution on training example i on round t (set of weights over the training examples). At each

Dr. Lu-yong, Wang is with Integrated Data Systems Department, Siemens Corporate Research, Princeton, NJ, 08540, (phone: 609-734-3671; fax: 609-734-6565; email: luyong.wang@siemens.com).

iteration t, the base learner is used to find a weak hypothesis h_t appropriate for the distribution. The weight will be updated. Usually, the weights of incorrectly classified examples are increased so that the base learner is forced on the hard examples in the training set. The base learner is called again with new weights over the training examples and the process iterates. At last, all the weak hypotheses are combined into a single strong hypothesis using a weighted majority vote (Please refer to details in Figure 1). The discriminative model corresponding to the string classifier, H(x), is

$$q(y|x) = \frac{e^{2yH(x)}}{1+e^{2yH(x)}}$$

The error rate ϵ is proven to be bounded by

$$\epsilon \leq 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)}$$

Input Samples:
$S = \{(x_1, y_1, w_i), ..., (x_n, y_n, w_i)\}$,where $x_i \in X, y_i \in Y : \{-1, +1\}$
Initialize $D_1(i) = \frac{w_i}{\sum w_i}$
For each $t = 1, ..., T$

(1) Train the base learner using distribution D_t
(2) Get Weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with error
$\epsilon_t = Pr_{i\sim W_t}[h_t(x_i) \neq y_i]$
(3) Choose $\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$
(4) Update:

$$D_{t+1}(i) = \frac{D_t(i)e^{(-\alpha_t y_i h_t(x_i))}}{Z_t} \quad (1)$$

where Z_t is a normalization factor
Output: Final hypothesis

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \quad (2)$$

Figure 1. AdaBoost Algorithm

One of the key features of AdaBoost is that misclassified samples in the previous training received more weights in the next time. However, AdaBoost can not rule out the chance that the correctly classified samples be misclassified again. Thus, this new probabilistic boosting tree (*PBTree*) algorithm utilized a divide-and-conquer approach in the training.

- *Probabilistic Boosting Tree*

In order to show our new method in a simple format, we denote the probabilities computed by each learned AdaBoost method as

$$q(+1|x) = \frac{e^{2H(x)}}{1+e^{2H(x)}}, q(-1|x) = \frac{e^{-2H(x)}}{1+e^{-2H(x)}}$$

As shown in Figure 2, *PBTree* learns a tree in the training process: at each node, a strong classifier is learned using the AdaBoost algorithm. The training samples are then divided into two new sets using the learned classifier: the left one and the right one. Then, a left sub-tree and a right sub-tree were trained respectively. To control the overfitting problem to a certain degree, variable ϵ is defined to show the support vectors, which means samples falls in the range of $[1/2 - \epsilon, 1/2 + \epsilon]$ are treated as confusing ones (support vectors), and they will be used in the left and right sub-trees for learning.

Input: Samples for training: $S = \{(x_1, y_1, w_1), ..., (x_n, y_n, w_n)\}$,
where $x_i \in X, y_i \in Y : \{-1, +1\}, \sum_i w_i = 1$;
(1) Compute the empirical distribution $\hat{q}(y) = \sum_i w_i \delta(y_i = y)$
(2) For training set S, train a strong classifier using AdaBoost algorithm with T weak classifier
(3) If $\epsilon_t < \theta$ (θ is a predefined error rate threshold), exit from the iteration
(4) If current tree depth is L, then exit
(5) Initialize two empty set S_{left} and S_{right}
(6) For each sample (x_i, y_i), compute the probability $q(+1|x)$ and $q(+1|x)$ using the learned strong classifier.
If $q(+1|x) - \frac{1}{2} < \varepsilon$, Then $(x_i, y_i, 1) \rightarrow S_{right}$
else if $q(-1|x) - \frac{1}{2} < \varepsilon$, en $(x_i, y_i, 1) \rightarrow S_{left}$
else $(x_i, y_i, q(+1|x_i)) \rightarrow S_{right}$; $(x_i, y_i, q(-1|x_i)) \rightarrow S_{left}$
(7) Normalized all the weights of the samples in the S_{left}
(8) Repeat the procedure
(9) Normalized all the weights of the samples in the S_{right}
(10) Repeat the procedure

Figure 2. Two-class Probabilistic Boosting Tree Training Algorithm

Function F_N(x,y) to compute posterior distribution $\hat{p}(y|x)$ at tree node N

1. For a given sample x, calculate $q_N(+1|x)$ and $q_N(-1|x)$ using the learned AdaBoost model at the current tree node N
2. If $q_N(+1|x) - \frac{1}{2} > \varepsilon$, then $\hat{p}_{right}(y) = F_{right(N)}(x, y)$ and $\hat{p}_{left}(y) = \hat{q}_{left(N)}(y)$ where $\hat{q}_{left(N)}(y)$ is the empirical distribution of the left tree.

   else if $q_N(+1|x) - \frac{1}{2} > \varepsilon$, then $\hat{p}_{right}(y) = \hat{q}_{right(N)}(y)$ and $\hat{p}_{left}(y) = F_{left(N)}(x, y)$

   else $\hat{p}_{right}(y) = F_{right(N)}(x, y)$ and $\hat{p}_{left}(y) = F_{left(N)}(x, y)$
3. $\hat{p}_N(y|x) = q(+1|x)\hat{p}_{right}(y) + q(-1|x)\hat{p}_{left}(y)$

Figure 3. Two-class Probabilistic Boosting Tree Testing Algorithm

In a similar way, the testing process for probabilistic boosting tree in a top-down fashion. As Figure 3 illustrates, the testing process begin from the top node. It gathers the information from its descendant and report an overall approximated posterior distribution. This algorithm can also turn into a classifier that makes hard decision. As $q(+1|x)$ and $q(-1|x)$ are calculated, one can decide to go into the right or left sub-trees by comparing these two probabilities. The prediction of y is made at the leaf node of the tree by checking the empirical distribution. Prediction result is then passed back to the top node of the tree.

B. Simulation and validation

For the purpose of intuitive understanding, a two-dimensional point distribution was simulated, which consists of a synthetic

dataset of 1,553 points. It is composed of 6 positive subgroups (total 573 points) and 7 negative groups (total 980 points). The simulation parameter settings are shown in Table 1, and the resulting point distribution (as well as its *PBTree* classification process) was shown in Figure 4.

Table 1. Simulation parameter settings in a two-dimension point distribution

class	cluster #	points	x				y			
			distrib	Mean	σ/w*	interval	distrib	Mean	σ/w*	interval
+	1	100	Gaussian	30	20	0.5	Uniform	70	105	0.5
	2	110	Gaussian	95	5	0.5	Gaussian	40	15	0.5
	3	100	Gaussian	0	10	0.5	Gaussian	40	10	0.5
	4	100	Uniform	60	92	0.5	Gaussian	0	12	0.5
	5	93	Gaussian	55	12	0.5	Gaussian	-40	12	0.5
	6	70	Gaussian	-70	12	0.5	Gaussian	50	20	0.5
-	7	50	Gaussian	80	10	0.5	Gaussian	90	20	0.5
	8	80	Uniform	20	50	0.5	Gaussian	15	20	0.5
	9	100	Gaussian	60	12	0.5	Gaussian	42	15	0.5
	10	110	Gaussian	-27	13	0.5	Gaussian	65	15	0.5
	11	70	Gaussian	0	13	0.5	Gaussian	0	9	0.5
	12	500	Gaussian	150	13	0.5	Gaussian	-20	25	0.5
	13	70	Gaussian	-30	13	0.5	Gaussian	-15	15	0.5

* σ refer to standard variance in Gaussian distribution; while w refer to width in uniform distribution

In high dimensional problems, many datasets can also be transformed into intuitive two-dimensional views by principle component analysis (PCA).

C. *Cancer Diagnosis based on gene expression profiling*

Lung adenocarcinomas are the most common form in lung tumor. The histopathological classification of lung adenocarcinoma is challenging. Also, metastases of non-lung origin are hard to distinguish from lung adenocarcinomas. More fundamental knowledge of molecular classification of lung adenocarcinomas may aid in the prediction of patient outcome and the informed selection of novel therapeutic target.

A total of 203 lung adenocarcinomas and its pathological control specimen were used to create 2 datasets. Total RNA extracted from samples was used to generate cRNA target, subsequently hybridized to human U95A oligonucleotide probe arrays. A standard deviation threshold of 50 expression units was used to select the 3,312 most variable transcript sequences. Among them, 139 are lung adenocarcinomous (127 cases are lung adenocarcinomas, 12 cases were adenocarcinomas suspected to be extrapulmonary metastases based on clinical history), and 64 pathological controls [15].

Our aim is to perform and evaluate the classification between lung adenocarcinomas and its pathological controls and concurrent subtype (extrapulmonary metastases origin) discovery for the lung adenocarcinomas using the gene expression profiles.

III. Results and Discussions

A. *Probabilistic c boosting tree formation on synthetic dataset: a joint classification and clustering process*

Under this model, positive and negative samples in the complex simulation dataset are naturally separated and divided into groups by discriminative learning. Figure 4 shows how the tree is learned and training samples are divided. Samples which are hard to classify are passed further down leading to the expansion of the tree. Clustering of positives and negatives is naturally performed by serving the other as auxiliary variables. Since each tree node is a strong classifier, it is capable of dealing with samples of complex distribution. There is no need to pre-specify the number clusters. Also, the hierarchical structure of the tree allows us to report the clusters according to different levels of discrimination.

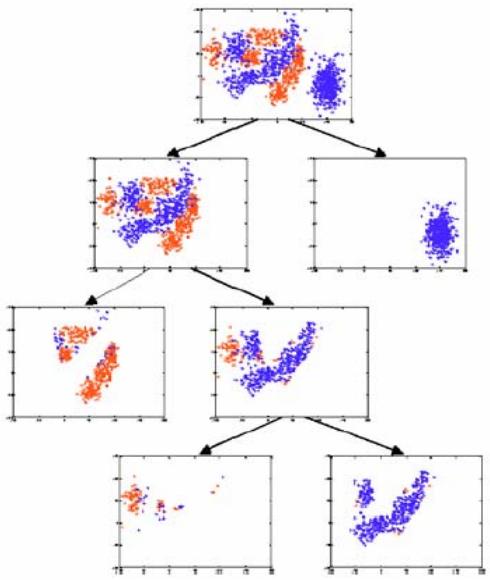

Figure 4. Probabilistic boosting tree on 2D synthetic data has clear classification and clustering capability.

We illustrate a probabilistic boosting tree on a synthetic dataset of 1,553 points. Weak classifiers are likelihood classifiers on features such as position and distance to line functions. The first level of the tree divides the whole set into two parts. The right side mostly has blue (dark) points since they are away from the rest of the clouds. The tree expands on the parts where positive and negative samples are tangled. Additionally, leave-one-out cross validation shows that the classification accuracy is around 98%, indicating that these methods has equivalently outstanding classification capacity with other advanced methods, such as Random Forest (leave-one-out cross validation accuracy is also around 98%).

B. *Lung Cancer Adenocarcinomas Diagnosis and subtype inference by probabilistic boosting tree*

We evaluated the predictive capability of this probabilistic tree using lung cancer dataset, which consists of 3,312 transcript variables. We aim to estimate the performance of probabilistic boosting on high dimensional gene expression data. The leave-one-out cross valuation (LOOCV) shows the classification accuracy is 88.7%, with the specificity of 86% and sensitivity of 90%. The results indicate strong classification capability of our PBtree algorithm, considering random forest, one of the best classification methods, achieves 86.2% accuracy. For resulting two-layer probabilistic boosting

tree, the training error is 0.0049, while the training error of resulting 3-layer probabilistic boosting tree, the training error reaches 0. Empirical p-value was estimated by permuting the dataset and re-evaluating the error rate. We permuting the dataset for 5000 times, none has lower error rate. It shows the p-value is below 1/5000.

	Positive	Negative
True	TP=125	TN=55
False	FP=9	FN=14
Error rate = 11.3%	Specificity = 0.86	Sensitivity = 0.90

Table 1 Leave-one-out-cross-validation using lung adenocarcinomas data by probabilistic boosting tree.

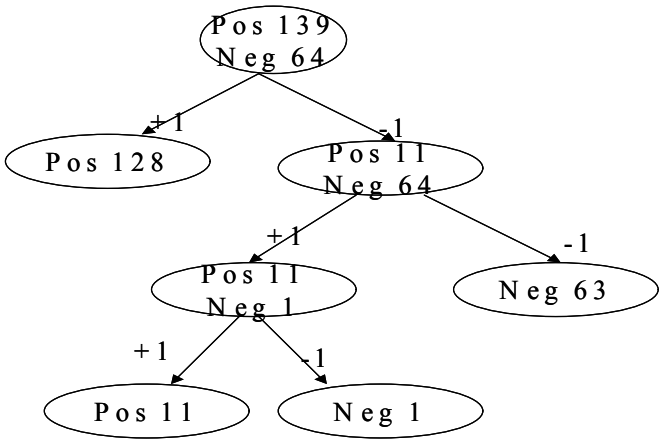

Figure 5. Probabilistic boosting tree training using lung adenocarcinomas

The resulting *PBtree* based on training was shown in Figure 5. It has clear demonstrated that the 11 out of 12 cases that were suspected to be extrapulmonary metastases among the adenocarcinomas cases were clearly separated out from other 128 adenocarcinomas cases. The subtype discovery process is coupled with discriminative training of probabilistic boosting tree, which is an inherent feature for this *PBTree* algorithm. It indicated that the *PBtree* can identify and separate the cases with extrapulmonary metastases, which is hardly discern from pathological examination. Thus, the *PBTree* learning and classification results based on both simulation dataset and real cancer gene expression data indicated that it can naturally incorporate excellent discriminative ability and subtype discovery inference in its inherent mechanism.

IV. Conclusions

We propose a novel disease diagnostic method based on probabilistic boosting tree algorithm, which can jointly classify and cluster data during its classification process. In the learning stage, the probabilistic boosting tree automatically constructs a tree in which each node combines a number of weak classifiers into a strong classifier. In the testing stage, the conditional probability is calculated at each tree node based on the learned classifier, which guide the probability propagation in its subtrees. Additionally, clustering is naturally embedded in the learning phase and each subtree represents a cluster of certain level. We show the probabilistic boosting tree algorithm and its evaluation in both simulation and real dataset.

The novelty of the above work is derived from the following: To our best knowledge, this is the pioneering method for joint classification and subtype discovery on expression profiles based on *PBTree* algorithm. It naturally incorporates a subtype discovery process, which is based on discriminative features without the need for pre-specify the cluster numbers. This method provides a powerful tool for doctors to make diagnosis and detect new subtype during the diagnostic process. This method is capable of extending to new domain, such as clinical decision support. Thus, it is capable of providing a useful tool in the later personalized medicine based on genetic profiles. This method is scalable for larger dataset, and may be utilized in database-guided diagnosis and information-based medicine.

References

[1] T. R. Golub, Slonim,D.K., Tamayo,P., Huard,C., Gaaasenbeek, M., Mesirov,J.P., Coller,H, Loh,M.L.,Downing,J.R., Caligiuri,M.A., et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 2002.

[2] S. Ramaswamy, Tamayo,P., Rifkin,R., Mukhejee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P., et al, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc Natl. Acad. Sci. USA*, vol. 98, pp. 15149-15154, 2001.

[3] S. Dudoit, Fridlyand, J. and Speed, T. P., "Comparison of discrimination methods for the classification of tumors using gene expression data," *Technical Report 576, Department of Statistics, University of California, Berkeley.*, 2000.

[4] D. V. a. R. Nguyen, D. M., "Tumor classification by partial least squares using microarray gene expression data.," *Bioinformatics*, vol. 18, pp. 39-50, 2002.

[5] A. Ben-Dor, Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, N., "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, pp. 559-584, 2000.

[6] N. Friedman, Linial, M., Nachman,I., Pe'er D., "Using Bayesian Network to Analyze Expression Data," *Journal of Computational Biology*, vol. 7, pp. 601-620, 2000.

[7] T. S. Furey, Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D., "Support vector machine classification and validation of cancer tissue samples using microarray expression data.," *Bioinformatics*, vol. 16, pp. 906-914, 2000.

[8] M. Dettling, Buhlmann,P., "Boosting for tumor classification with gene expression data," *Bioinformatics*, vol. 19, pp. 1061-1069, 2003.

[9] U. Alon, Notterman, D.A., Gish,K., Ybarra,S., Mack,D., and Levine, A.J., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by olligonuceotide arrays," *Proc Natl. Acad. Sci. USA*, vol. 96, pp. 6745-6750, 1999.

[10] Y. Liu, Ringer M., "Multiclass discovery in array data.," *BMC Bioinformatics*, vol. 5, pp. 70, 2004.

[11] S. Varma S., R., "Iterative class discovery and feature selection using minimal spanning trees.," *BMC Bioinformatics*, vol. 5, pp. 126, 2004.

[12] R. Alexandridis, Lin, S., and Irwin M., "Class discovery and classification of Tumor samples using mixture modelling of gene expression data," *Bioinformatics*, pp. Advanced Access, April 29, 2004, 2004.

[13] T. R. Golub, et al, "Molecular classification of Cancer: Class discover and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[14] Y. a. S. Freund, RE, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, pp. 771-780, 1999.

[15] A. Bhattacharjee, et al, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc Natl. Acad. Sci. USA*, vol. 98, pp. 13790-13795, 2001.