Diagnosis and Biomarker Identification on SELDI proteomics data by ADTBoost

Lu-yong Wang* Amit Chakraborty Dorin Comaniciu
Integrated Data Systems Department
Siemens Corporate Research
755 College Road East, Princeton, NJ, 08540
Luyong.Wang@siemens.com

Abstract

Clinical proteomics is an emerging field that will have great impact on molecular diagnosis, identification of disease biomarkers, drug discovery and clinical trials in the post-genomic era. Protein profiling in tissues and fluids in disease and pathological control and other proteomics techniques will play an important role in molecular diagnosis with therapeutics and personalized healthcare. We introduced a new robust diagnostic method based on ADTboost algorithm, a novel method in proteomics data analysis to improve classification accuracy. It generates classification rules, which are often smaller and easier to interpret. This method often gives most discriminative features, which can be utilized as biomarkers for diagnostic purpose. Also, it has a nice feature of providing a measure of prediction confidence. We carried out this method in Amyotrophic lateral sclerosis disease data acquired by surface enhanced laser desorption/ionization-time-of-flight mass spectrometry experiments. Our method is shown to have outstanding prediction capacity through the cross-validation, ROC analysis results and comparative study. Our molecular diagnosis method provides an efficient way to distinguish ALS disease from neurological controls. The results are expressed in a simple and straightforward alternating decision tree format or conditional format. We identified most discriminative peaks in proteomic data, which can be utilized as biomarkers for diagnosis. ADTboost is not only useful in on proteomic data classification, it can also integrate other clinical, imaging data from heterogeneous sources for early diagnosis. It will have broad application in molecular diagnosis through proteomics and personalized medicine.

1. Introduction

Proteomics techniques have made large-scale determination of gene and cellular function at protein level feasible in this post-genomic era. It has become an important screening tool for the discovery of potential biomarkers and molecular diagnosis. Surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) mass spectrometry provides a sensitive system to detect and resolve the multiple proteins bound to protein chip arrays [4]. It has gained success in biomarker discovery for different types of cancers [1] [2] [4]. Recent reports have raised the expectation for the application of proteomic profiling to clinical diagnosis [1].
The analysis and interpretation of the enormous volumes of proteomic data remains an unsolved bioinformatics challenge despite the recent advances in proteomic technology. Few approaches have been reported for classification problem on SELDI-TOF mass spectrometry data. These approaches include Fisher discriminative analysis, CART, Support Vector Machine,non-parametric kernels, nearest neighbor methods,boosted decision stump [5] and genetic algorithm [6] [7]. In statistical learning, ADTboost is a supervised classification algorithm, which improves and generalizes decision trees, boosted decision stumps and boosted decision tree in a natural way [8]. ADTboost has been proved in statistical learning that it provides significant improvement in classification error than single decision trees. Its performance is similar to C5.0 with boosting[8](C5.0 program is available only commercially from Rulequest Research). ADTboost generates rules that are relatively smaller in size and easier to interpret [8].
In this study, we introduce an automatic molecular diagnostic method based on ADTboost to classify SELDI-TOF data and identify most discriminative peak biomarkers (m/z ratios) between Amyotrophic lateral sclerosis (ALS) disease and neurological control SELDI-TOF dataset.

2 ADTBoost and ALS SELDI MS diagnosis

Here we describe very briefly the ADTBoost algorithm: We use AdaBoost to learn decision rules constituting alternating decision tree and combining these rules through a weighted majority voting . Its input is $(x_1, y_1), \cdots, (x_n, y_n)$, where x_i is a vector of protein profile in SELDI data analy-

sis, and y_i belongs to label set Y $\{+1, -1\}$. Its output is in a form of alternating decision tree for classification([8]). The resultant alternating decision tree contains splitter nodes (associated with a test) and prediction nodes (associated with a value). Each prediction node represents a weak prediction rule. At each boosting iteration, a new splitter node with its prediction nodes is introduced. The classification associated with an instance is the sign of the sum of the predictions along the related paths in the tree.

We carried out our method on SELDI MS (surface enhanced laser desorption ionization mass spectrometry) data of 36 ALS patients and 31 neurological controls acquired in ALS disease research from our collaborating lab to diagnose ALS disease based on protein expression profiles in CSF and identify specific sets of ALS diagnostic protein fingerprint in CSF through SELDI-TOF experiments to be used in the clinical setting.

To evaluate diagnostic capacity of ADTboost, we carried out ten fold cross validation using SELDI MS data acquired in ALS disease research. We found that the sensitivity of this ALS diagnostic method for ALS disease reaches 77.8% and the specificity reaches 77.4%. On average, 77.6% of the specimen were correctly classified.

ROC curves in Fig 1 show the false positive rate and true positive rate for our diagnostic method with respect to different numbers of boosting training iteration. It shows that as the more training rounds get more discriminative peak values for the classification, the more accurate diagnostic result may be achieved. Meanwhile, it shows that the ROC curve of classification by 5 boosting training rounds (top four significant SELDI peaks's m/z ratio value) in this dataset is near the upper ROC curve of classification by 30 boosting training rounds. ADTboost aids to identify most valuable biomarkers in a top-down manner on splitter nodes in the alternating decision tree.

As expected, since ALS disease is a complex neurodegenerative disease that is not so simply associated with genetic causes as cancer, the sensitivity and specificity of ALS diagnosis can not be as ideal as 95% and 94% in cancer. However, the results of cross validation indicated that ALS disease is also capable of molecular diagnosis and biomarker identification.

We compare the performance of this method with the prior methods. It shows that ADTboost method perform well among the other methods in classification error estimation on ALS data set. These traditional methods are linear discriminate analysis (LDA), k-nearest neighbour(kNN),support vector machine (SVM), Random Forest(RF). It has the lowest error rate among these classifiers. LDA is the second lowest. RF and SVM has close performance to the top two methods. kNN methods perform the worst.

Based on our evaluation on ALS disease dataset, we

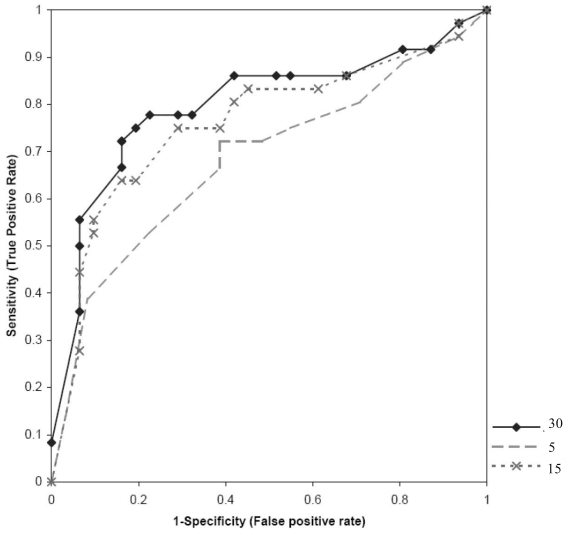

Figure 1. ROC curves based on ADTboost ALS classification

can conclude that ADTboost is capable of diagnosis using SELDI-TOF on tissue samples and capable of providing informative biomarkers for diseases. It provides an intuitive and outstanding diagnostic method for clinical practice. It can also extend and integrate with other clinical,imaging data from heterogeneous sources to improve accurate and early diagnosis.

References

[1] R. Aerbersold. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, March 2003.

[2] H. Kuruma. Proteome analysis of prostate cancer. *Prostate Cancer and Prostatic disease*, 8(1):1–8, March 2004.

[3] Z. Lin. Application of seldi-tof mass spectrometry for the identification of differentially expressed proteins in transformed follicular lymphoma. *Modern Pathology*, 17(6):670–678, June 2004.

[4] E. Petrocoin. Use of proteomic patterns in serum to identify ovian cancer. *Lancet*, 259(9306):572–577, Feb 2002.

[5] Y. Qu. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48(10):1835–1843, October 2002.

[6] M. Wagner. Computational protein biomarker prediction: a case study for prostate cancer. *BMC bbioinformatics*, 5(1):26–35, March 2004.

[7] C. White. Bioinformatics strategies for protein profilings. *Clinical Biochemistry*, 37(7):636–641, July 2004.

[8] F. Yoav. The alternating decision tree learning algorithms. *Proceedings of 16th International Conference on Machine Learning*.