

Boosting Methods in Classification and Analysis on mass spectrometry proteomic data

Lu-yong Wang*, Amit Chakraborty, Dorin Comaniciu ¹

Keywords: mass spectrometry, boosting, alternating decision tree, SELDI-TOF

1 Introduction.

Recent reports have raised the expectation for the application of proteomic profiling to clinical diagnosis. ([1],[3],[6]). It becomes a crucial component for emerging personalized medicine. Although several methods, *e.g.* FDA, SVM, CART, non-parametric kernels, kNN, boosted decision stump and genetic algorithm, have been reported for class discrimination on mass spectrometry data([4],[5]), it remains an unsolved challenge to analyze and interpret the enormous volumes of proteomic data. Here, we introduce ADTBoost algorithm to SELDI-TOF data analysis. ADTBoost improves and generalizes decision trees, boosted decision stumps and boosted decision tree in a natural way. ADTBoost provides significant improvement in classification error over single decision trees. Its performance is similar to C5.0 with boosting (C5.0 is available commercially from Rulequest Research). ADTBoost generates simpler and intuitive rules, and yields a natural measurement of classification confidence([2], [7]). We show by example that ADTBoost can not only classify the protein profiles between patients and controls, but also aid to identify most discriminative biomarkers between disease and normal on SELDI-TOF MS dataset.

2 Method.

Here we only 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 $(x_1, y_1), \dots, (x_n, y_n)$, where x_i is a vector of protein profile in SELDI data analysis, and y_i belongs to label set $Y \{+1, -1\}$. Its output is in a form of alternating decision tree for classification([2], [7]). 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.

3 Results and Conclusion.

To benchmark the prediction capacity of this method, we carried out 10-fold cross validation using SELDI MS peaks data with WCX ProteinChips acquired in ALS disease research from our collaborating lab from Mount Sinai School of Medicine. We applied our method on SELDI MS data of Amyotrophic lateral sclerosis (ALS, a progressive neurodegenerative disease that affects neurons in brain and spinal cord) patients and neurological controls.

We evaluated the prediction capacity of ADTBoost by calculating the true positive (TP) hits, false positive (FP) hits, true negative (TN) hits, false negative (FN) hits, sensitivity

¹Integrated Data System Department, Siemens Corporate Research, 755 College Road East, Princeton, NJ, USA. *Correspondence author: E-mail: Luyong.Wang@siemens.com

and specificity. We found that the sensitivity of this ALS diagnostic method is 77.8% and the specificity is 77.4%. As expected, since ALS disease is a complex neurodegenerative disease, it is not a simply-genetic disease like cancer. The sensitivity and specificity of ALS diagnosis can not be as high as 95% and 94% in cancer. However, the results of cross validation indicated that ALS is also capable of molecular diagnosis and biomarker identification.

Also, based on the resultant alternating decision tree, we found that top significant SELDI peaks's m/z ratio value are 6690, 22516, 23211, 1185 in the top-down order. These peaks were selected by ADTBoost to minimize the training error at each step. ADTBoost aids to identify valuable biomarkers in a top-down manner on splitter nodes in the alternating decision tree. We compared the classification results with the our univariate differential analysis t-test for different peaks: The most important decision rules based on our ADTBoost results are consistent with univariate differential analysis.

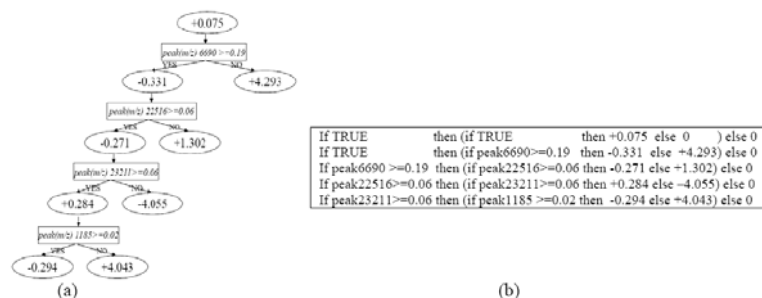


Figure 1: Example of classifier in an alternating-tree format by 5 round ADTBoost training; (a) alternating decision tree and (b) corresponding decision rules

In conclusion, we can conclude that ADTBoost is capable of diagnosis using SELDI-TOF on tissue samples and capable of providing informative biomarkers for diseases based on our evaluation on ALS disease dataset. The advances in proteomics will provide more insights in the clinical biochemistry and personalized healthcare, and ADTBoost will be useful in robust classification and biomarker identification.

References

- [1] Aebersold, R., Man, M. 2003 Mass spectrometry-based proteomics, *Nature*, 422: 198-207
- [2] Homes, G., Pfahringer, B., Kirkby, R., Frank, E., Hall, M., 2002, Multiclass alternating decision trees. *Proceedings of the European Conference on Machine Learning, Spring Verlag*
- [3] Kuruma, H, Egawa, S, Oh-Shi, M., Kodera, Y., Maeda, T. 2004, Proteome Analysis of prostate cancer, *Prostate Cancer and Prostatic disease*, 1: 1-8
- [4] Petrocino, EF, Ardekani, AM, Hitt, BA, Levin, PJ, Fusaro, V.A., Steinberg, SM, Mills, G.B., Simone, C, Fishman, DA, Kohn, EC, and Liotta, LA (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 259: 572-577
- [5] Wagner, M., Naik, D., Pothan, A., Kasukurti, S., Devineni, R.R., Adam, B.L., Semems, O.J. and Wright G.L.Jr. 2004 Computational protein biomarker prediction: a case study for prostate cancer, *BMC bioinformatics*, 5: 26-35
- [6] White, C.N., Chan, D.W., Zhang, Z. 2004 Bioinformatics strategies for protein profiling, *Clinical Biochemistry*, 37: 636-641
- [7] Yoav, F., Mason, L. 1999 The alternating decision tree learning algorithm, *Proceedings of 16th International Conference on Machine Learning*, 124-133.