Poster G-15

Modified Random Forest Algorithm for Biomarker Identification in Clustered Mass Spectrometry Data

Authors:
Yuliya Karpievitch (*Dept. of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center, University of Texas*)
Elizabeth Hill (*Dept. of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina*)
Timothy Millar (*Dept. of Gastroenterology and Hepatology, Medical University of South Carolina*)
Adam Smolka (*Dept. of Gastroenterology and Hepatology, Medical University of South Carolina*)
Jonas Almeida (*Dept. of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center, University of Texas*)
Brenda Hoffman (*Dept. of Gastroenterology and Hepatology, Medical University of South Carolina*)

Short Abstract: We present a novel modification of the Random Forest (Breiman 2001) algorithm for biomarker identification in clustered or longitudinal data commonly seen in biological experiments. Clustered data are not independent identically distributed (iid) and are typically positively correlated, thus we provided enhancements to accommodate for explicit relationships between input parameters.

Long Abstract:

We present a novel modification of the Random Forest algorithm to accommodate clustered data commonly seen in mass spectrometry and other biological experiments [1]. Random Forest builds an ensemble of trees constructed from independent identically distributed (iid) random vectors, i.e. one observation vector per subject. Each tree in the forest casts a vote for the most popular classification. Random Forest is a robust learning algorithm, but growing a Random Forest based on clustered data violates the iid assumption, as clustered data are typically positively correlated and calls into question the reliability of resulting classification and variable importance measures. Our addition to the original Random Forest algorithm contains enhancements to deal with correlated data for classification and disease biomarker identification. We apply our methods to matrix-assisted laser desorption/ionization (MALDI) data, with replicate spectrograms collected for each sample. MALDI is widely used to discover disease-related biomarkers from easily obtainable bodily fluids like urine, saliva or serum [2-5]. The heterogeneous nature of the crystallized samples spotted on the MALDI plate often requires multiple measurements from the same individual resulting in clustered data [6]. We analyzed MALDI esophageal cancer data obtained at the Medical University of South Carolina, as well as the virtual mass spectra data obtained form the 'virtual mass spectrometer' developed at the M.D. Anderson Cancer Center. The data were preprocessed using a peak detection algorithm based on the mean spectrum [7].

Our novel contribution to the Random Forest algorithm consists of four parts.
First, Random Forest grows each tree on a bootstrap sample (random sample selected with

replacement) of the training data. We modified the bootstrap algorithm to take samples at the subject level rather then at the spectral level, so that bootstrap samples are constructed from independent units with correlated subunits. Second, we introduced weights for each tree based on tree-level accuracy and consistency. We consider a tree to display accuracy if correct classification replicates for a particular subject is observed. We consider a tree to behave consistently if subjects are classified consistently even if that classification is incorrect. This allows trees that do best in one or both of these categories to have a larger weight in voting, thus strengthening the classification power of the predictor. Next, we devised an improved node splitting rule. This rule encourages a split that propagates all the replicates for a given subject down the same branch and towards the same terminal node of the tree. This modification can work as an alternative to the weights procedure previously described, or it can work in conjunction with the tree weights to provide a better predictor. Finally, Random Forest, provides an unbiased error rate based on out-of-bag (OOB) data. We added a running unbiased error estimate for each subject. This allows small misclassifications, for example, one or two out of ten replicas, not to affect the forest error rate. This is useful, because we are ultimately interested in subject-level classification rather then classification of individual replicates from the same subject.

[1] Breiman LE (2001) Random Forests. Machine learning 45:5-32.
[2] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. (2002) Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359: 572 – 577.
[3] Schaub S, Wilkins J, Weiler T, Sangster K, Rush D and Nickerson P (2004) Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. Kidney International 65:323-332.
[4] Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z and Wright GL (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Research 62:3609–3614.
[5] Adam PJ, Boyd R, Tyson KL, Fletcher GC, Stamps A, Hudson L, Poyser HR, Redpath N, Griffiths M, Steers G, et al. (2003) Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. J. Biological Chemistry 278:6482-6489.
[6] Garden RW, Sweedler JV (2000) Heterogeneity within MALDI samples as revealed by mass spectrometric imaging. Analytical chemistry 72:30-36.
[7] Morris JS, Coombes KR, Koomen J, Baggerly KA and Kobayashi R (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 21:1764-1775.