

FIMXII-SCMA2005@AUBURN, Twelfth Annual International Conference on Statistics, Combinatorics, Mathematics and Applications, December 2–4, 2005, Auburn University, Auburn, Alabama, USA

Coauthors: Lara M. DePadilla

**FEATURE SELECTION AND MACHINE LEARNING WITH
MASS SPECTROMETRY DATA FOR DISTINGUISHING
CANCER AND NON-CANCER SAMPLES**

SUSMITA DATTA

In this talk, I present a comparative study of various clustering and classification algorithms as applied to differentiate cancer and non-cancer protein samples using mass spectrometry data. Our study demonstrates the usefulness of a feature selection step prior to applying a machine learning tool. A natural and common choice of a feature selection tool is the collection of marginal p-values obtained from t-tests for testing the intensity differences at each m/z ratio in the cancer versus non-cancer samples. We study the effect of selecting a cutoff in terms of the overall Type 1 error rate control on the performance of the clustering and classification algorithms using the significant features. For the classification problem, we also considered m/z selection using the importance measures computed by the Random Forest algorithm of Breiman. Using a data set of proteomic analysis of serum from ovarian cancer patients and serum from cancer-free individuals in the Food and Drug Administration National Cancer Institute Clinical Proteomics Database, we undertake a comparative study of the net effect of the machine learning algorithm - feature selection tool - cutoff criteria combination on the performance as measured by an appropriate error rate measure.

Keywords: Mass spectrometry, high throughput data, clustering, classification, machine learning, microarray.

DEPARTMENT OF BIOINFORMATICS AND BIostatISTICS, UNIVERSITY OF LOUISVILLE
E-mail address: `susmita.datta@louisville.edu`