Biomarker stability of multiresolution proteomic profiling

ID - 173

Furlanello Cesare[1], Jurman Giuseppe[1], Riccadonna Samantha[1,2], Barla Annalisa[3], Paoli Silvano[1,2], Merler Stefano[1]

[1]IRST, FBK, Trento

[2]DIT, University of Trento, Trento

[3]DISI, University of Genova, Genova

Motivation

In functional genomics studies such as classification from microarray data, resampling-like procedures are accepted as a standard caution against selection bias. Replicas of the data (e.g. by bootstrap), combined with partitioning of the available data in training and test sets are in particular common in profiling studies with high-throughput data. In practice, control of bias and variance is achieved by paying the cost of having to deal with sets of alternative output solutions: biomarker lists may be fairly different on different versions of the data, turning the identification of the most important features into a less defined task. We propose to employ measures of stability for ranked list sets as a tool for assessing the reliability of molecular profiles [1]. In particular, we consider the stability of biomarker lists in experiments with different setups. Analysis of proteomic data from mass spectrometry (MS), as discussed in this paper, is especially affected by the dependence from preprocessing choices (e.g. in the peak extraction procedures). The list of biomarkers can vary by tuning parameters dependent on the location in the m/z range and on the physical characteristics of the MS system being used. In [2], a multiresolution technique was proposed to automate the proteomic preprocessing pipeline in predictive classification of MS cancer data. Here we consider a new type of stability measure to analyze multiresolution spectra profiling.

Methods

Two MS datasets were employed: a synthetic one generated with the Cromwell simulator [3] and an Ovarian Cancer dataset introduced in [4]. For both datasets the task is predictive binary classification and selection of a ranked list of discriminant biomarkers. Alternative models are obtained by different parameter values in preprocessing [5]: we vary peak width and peak gap parameters in the search of candidate peaks to be used as features for the classifier/ranking algorithm. The use of 10 alternative resolutions (R1 R10) provides different instances of the preprocessed dataset. All the retrieved candidate peaks are then used to build a multiresolution version (MR) of the dataset. The complete validation platform BioDCV is applied for developing Support Vector Machine (SVM) classifiers and ranked feature lists. In all experiments, a subset of the data is reserved as independent test set for further control on prediction accuracy (procedure repeated 10 times on the Cromwell data and 5 times on Ovarian data). Measures of accuracy (average expected test error: ATE) and of stability (union number Iu(k): number of different features in the top-k sublists in 100 replicated experiments) are then computed.

Results

In both the profiling tasks, the MR model results both for accuracy and stability are close to the best performing model obtained by manually defining a specific resolution. The stability indicator Iu(k) diagnostic plots show that the MR models select a smaller subset of features and optimal or suboptimal ATE. If we consider the number of resulting candidate features (peak centroids corresponding to peptide families), it is found that the MR strategy selects the same number of peaks of high importance (those extracted in 75% of the 100 lists) than the single resolution models, although starting from much higher number of candidates.

Availability: http://biodcv.itc.it

Image: http://mpa.itc.it/static/Jurman_et_al_BITS2007/diagPlot.jpg

Email: furlan@itc.it