A Strategy for Classifying Mass Spectra

Ceccarelli M[1, 2], d’Acierno A[3] Facchiano A[3]

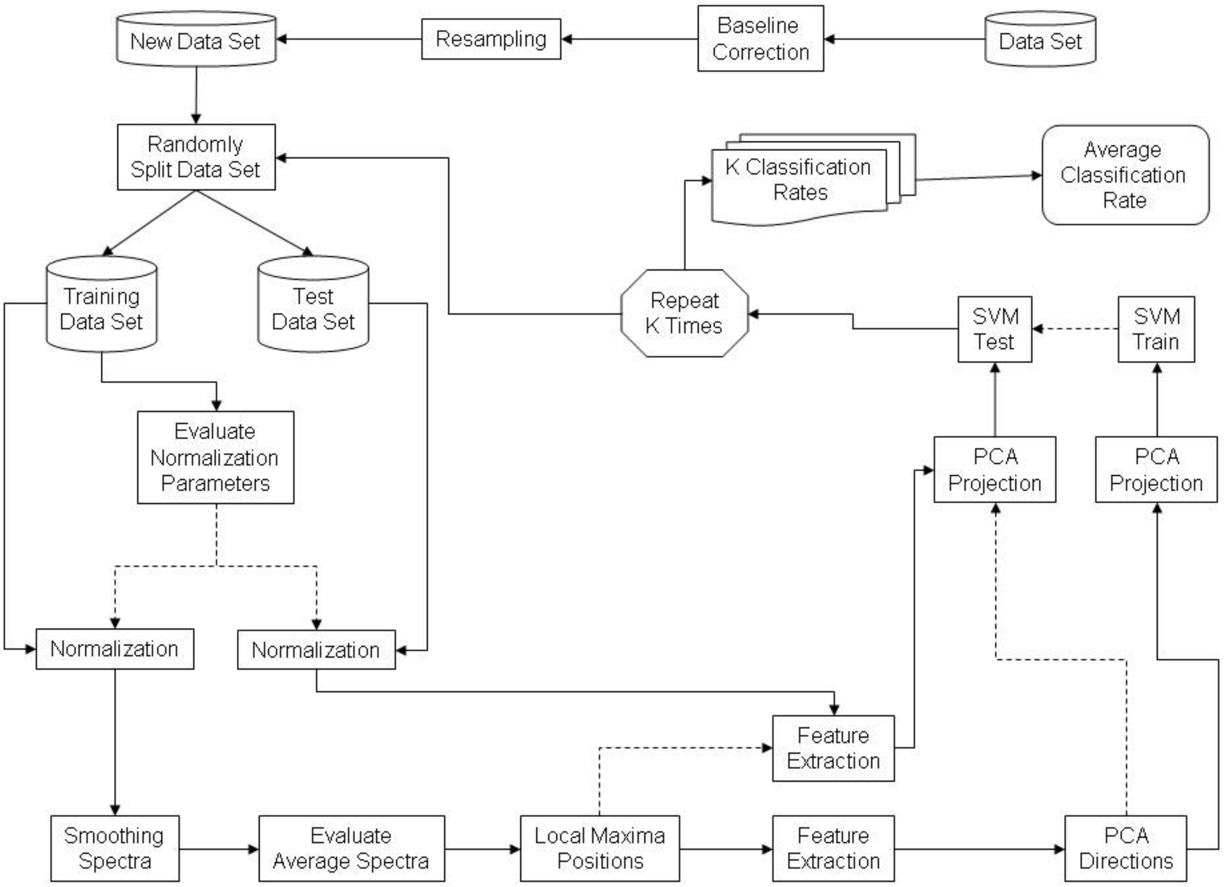
[1]Department of Biological and Environmental Sciences, University of Sannio,
Via Port'Arsa 11, 82100, Benevento, Italy
[2]Research Center on Software Technologies, University of Sannio,
Via Traiano 1, 82100, Benevento, Italy
[3]Institute of Food Sciences, National Research Council,
Via Roma 52 A/C, Avellino, Italy

Motivation

Mass spectrometry is the elective technique to characterize the proteome and its modification. The mass spectrum represents a molecular profile of the sample under analysis, obtained with increasing precision and automation techniques and, despite the large number of signals obtained in the proteome analysis, molecular modifications can be detected and markers of pathological states can be identified. In MALDI- and SELDI-TOF techniques proteins are co-crystallized with UV-absorbing compounds, then a UV laser beam is used to vaporize the crystals, and ionized proteins are then accelerated in an electric field. The analysis is then completed by the TOF analyzer. The spectra obtained can be profitably used for biomarkers discovery and other proteomic studies in biomedicine. In this paper we describe a system we have implemented for the automatic classification of SELDI spectra.

Methods

Data produced by mass spectrometry are spectra, typically reported as vectors of data, describing the intensity of signals due to biomolecules with specific mass-to-charge ratio values. Given the high dimensionality of spectra, given their different length and since they are often affected by errors and noise, preprocessing techniques are mandatory before any data analysis. After preprocessing (to correct noise and reduce dimensionality), several statistical and artificial intelligence based technologies could be used for mining these data. Figure 1 shows the overall process used to test our solution. After having independently corrected the baseline and re-sampled each spectra, we started k-fold cross validation (we used k = 10). As it is well known, in k-fold cross validation the data set is randomly divided in k sets; of the k sets, a set is retained as the validation data for testing the model, and the remaining k – 1 subsamples are used as training data. The cross-validation process is repeated k times, with each of the k subsets of samples used exactly once as the validation data. The k results from the folds are then averaged to produce a single estimation. Using the training set we derive the normalization parameters that are used to normalize both the training and the test sets. The normalized training data set is then used for feature extraction obtaining the m/z's of the peaks that best describe (according our method) each spectrum; these m/z's are then used to synthetically represent both the training spectra and the test spectra. Then, the training set is used to obtain PCA directions (obtained having fixed the overall energy); these directions are of course used to project both the training and the test sets. Last, the training set is used to train our SVM while the test set is clearly used to test the correct classification rate. It is worth noting that we perform the feature selection step externally with respect to the cross validation procedure; when the feature selection is done by using all the data and the performance evaluation by cross validation is performed just for the classification phase, in fact, then the obtained results may be severely biased due to the so called selection bias effect.



Results

We have tested our system using a well known dataset available from the National Cancer Institute of the U.S. National Institutes of Health consisting of 121 cancer samples and 95 control samples. In order to get the results all the classification experiments and estimated classification rates are averaged over 1000 runs of the whole process. The classification accuracies on these runs clusters around the average just like a Gaussian shape with average 0.9818 and standard deviation 4.314 * 10-5. As we can notice the classification accuracy is almost stable over the runs even if each run could eventually extract different peak sets to perform the classification. Moreover, we obtain a 100% accuracy in some of the runs but without using discrimination based peak selection.

Contact : antonio.dacierno@gmail.com