Neural Networks for Mass Spectra Classification: Preliminary Results

M. CECCARELLI\textsuperscript{1}, A. D’ACIERNO\textsuperscript{2}, A. FACCHIANO\textsuperscript{2}, A. MARATEA\textsuperscript{1}

\textsuperscript{1}RCOST, University of Sannio, viale Traiano 1, 82100 Benevento, Italy
Email: \{ceccarelli, amaratea\}@unisannio.it
\textsuperscript{2}ISA-CNR, via Roma 52, Avellino, 83100, Italy
Email: \{antonio.dacierno, angelo.facchiano\}@isa.cnr.it

Abstract: In this paper we summarize some experiments we have carried out in order to classify mass spectra. First we describe the pre-processing and the features’ extraction we have realized using the Bioinformatics toolbox of MATLAB. Then we describe results we have obtained using feed forward neural networks trained with a classical back-propagation; to avoid reinventing the wheel, WEKA has been used for this task. Preliminary results obtained by the analysis of a data set of tumor/healthy samples allowed us to correctly classify more than 80\% of samples. A more detailed analysis of our results show that our method at least deserves attention and further studies.

Keywords: Mass Spectra, Supervised Neural Nets, MATLAB, WEKA.

1. Introduction

Mass spectrometric data was optimized for high throughput comparison of samples that allows discovery, for example using serum profiling, of diseases’ biomarkers [1]. Data produced by mass spectrometry (the spectra) are represented by a (typically) very large set of measures representing the quantity of biomolecules having specific mass-to-charge (m/z) 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 artificial intelligence based technologies could be used for mining these data. In [1], for example, genetic algorithms and self organising maps were used to distinguish between healthy women and those affected by ovarian cancer. Support Vector Machines (combined with Particle Swarm Optimization) have been used [8] to distinguish cancer patients from non-cancer controls. Principal Component Analysis has been used for dimensionality reduction followed by linear discriminant analysis on SELDI spectra of human blood serum [2]. Several statistical methods have been compared in [3]. Classification using nearest centroid classification has been used in several applications; in [3], for example, it is used for protein mass spectrometry while ant colony optimization has been used [5] for peak selection from MALDI-TOF spectra.

In this paper we describe preliminary results we have obtained in some experiments performed having in mind to use off the shelf tools so that no new complex software has to be developed. Thus, we used MATLAB [6] and the Bioinformatics Toolbox for preprocessing data and WEKA [7] for the classification step.

Starting from a dataset of spectra we first pre-processed the data to obtain a set of comparable samples, then we extracted a set of fixed length features based on the most relevant peaks from spectra and finally applied a feed forward neural network trained with...
the classical back-propagation algorithm to predict the class of belonging of each sample. On a data set of tumor/healthy spectra we correctly classify (on average) more than 80% of samples as will be described in detail in the Results and Discussion section.

2. Methods

We have organised our two-classes spectra into sets of text files. Then, since we use MATLAB to pre-process our data, we read rough spectra using *dlmread* obtaining two cell arrays (one for each class); so we are in the position of easily pre-process our data. For each spectrum we adjusted the variable baseline using *msbackupd* that performs three steps: estimates the baseline within multiple shifted windows of width 200 m/z, regresses the varying baseline to the window points using a spline approximation and adjusts the baseline of the spectrum.

To make data easily comparable, we re-sampled each spectrum on 1000 samples using *msresample*: re-sampling homogenizes the mass/charge (M/Z) vector, allowing comparing different spectra under the same reference and at the same resolution. Moreover, the large size of the files leads to computationally intensive algorithms while high-resolution spectra can be redundant. By re-sampling, we decimate the signal into a more manageable vector, hopefully preserving the information content of the spectra.

Last, since in repeated experiments it is common to find systematic differences in the total amount of desorbed and ionized proteins we used the function *msnorm* that normalizes a group of mass spectra by standardizing the area under the curve to the group median.

Once spectra have been pre-processed we have to extract some significant features; in this paper we used peaks’ intensities and thus we uses *mspeaks* to extract the peaks of each spectrum; *mspeaks* finds peaks by first smoothing the signal using un-decimated wavelet transform with Daubechies coefficients, then assigning peak locations, and lastly, eliminating peaks that do not satisfy specified criteria. Again for each spectrum peaks are ordered according their intensity and the first 50 are selected. So, a dataset composed of $N$ spectra, resulted in a 50x$N$ matrix of peaks; each row clearly belonging to a well defined class. Last, this matrix has been converted in a *arff* file that can be easily used for classification by WEKA.

The architecture of the neural net we used, a feed forward one, is characterized by two output neurons (one for each class), clearly there are as many input neurons as the number of selected peaks and the network has just 1 hidden layer.

3. Results and Discussion

The data set we used [1] is composed by spectra derived from analysis of human serum from 95 unaffected women and 121 with ovarian cancer. After preprocessing and features extraction, a feed forward neural network, trained using the well known momentum back-propagation learning algorithm, has been used for classification. Inputs (i.e. peaks’ intensity) are normalized in the range (-1,1).

Finally, *k*-fold (with $k$ =10) cross validation has been used to test the generalization performance; here, as it is well known, the data set is divided in $k$ subsets and $k$ trials are performed using one the subsets as test sets and the remaining $k-1$ subsets as training set. Since the performance of our classifier of course depends on (at least) two parameters: the number of hidden neurons and the number of peaks to be used (i.e. the number of input neurons) we have performed several tests. Each net has been trained for 5000 epochs and each run has been repeated 10 times. Results are described in figure 1 where we show, for each pair (Peaks, Hidden Neurons) we have tested, the minimum, the maximum and the average percentage of correct classification on the test set. The standard deviation is also shown.
The neural networks perform quite well since, for each tested pair, the average correct classification rate is greater than 80% while a correct classification of at least 90% has been obtained (for each tested pair) in at least the 15% of trials; a correct classification lower than 70% has been obtained in at most the 10% of trials.

Overall, the proposed approach, in the authors’ opinion, certainly deserves attention and further studies for future improvements and testing.

Figure 1. Experimental results the number of peaks and the number of hidden neurons vary.

Acknowledgments

This work has been partially supported by the CNR project “Bioinformatics”.

4. References