A MACHINE LEARNING APPROACH TO MASS SPECTRA CLASSIFICATION WITH UNSUPERVISED FEATURE SELECTION

Michele Ceccarelli[(1)], Antonio d'Acierno[(2)], Angelo Facchiano[(2)]

(1) RCOST - University of Sannio
viale Traiano 1 - 82100 Benevento - Italy, ceccarelli@unisannio.it

(2) ISA-CNR
via Roma 52 - Avellino - 83100 - Italy, {antonio.dacierno, angelo.facchiano}@isa.cnr.it

Keywords: mass spectra, support vector machine, classification.

Abstract. Mass spectrometry spectra are recognized as a screening tool for detecting discriminatory protein patterns. Mass spectra, however, are high dimensional data and a large number of local maxima (a.k.a. *peaks*) have to be analyzed; to tackle this problem we have developed a three-step strategy. After data pre-processing we perform an unsupervised feature selection phase aimed at detecting salient parts of the spectra which could be useful for the subsequent classification phase. The main contribution of the paper is the development of this feature selection and extraction procedure grounded on the theory of multi-scale spaces. Then we use support vector machines for classification. Results obtained by the analysis of a data set of tumor/healthy samples allowed us to correctly classify more than 95% of samples. ROC analysis has been also performed.

1 Introduction

SELDI-TOF (Surface-Enhanced Laser Desorption and Ionization Time-Of-Flight) technology is considered a modified form of MALDI-TOF (Matrix-Assisted Laser Desorption and Ionization Time-Of-Flight). According to these 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. Differences in the two technologies, which reside mainly in the sample preparation, make SELDI-TOF more reliable for biomarkers discovery and other proteomic studies in biomedicine. The proteomic characterization by means of TOF (both SELDI and MALDI) technologies of samples from individuals is considered to carry information about the healthy or pathological state of the individual. In fact, samples as serum, plasma, and other kinds of extracts contain proteins for which the covalent structure may be modified in specific pathologies, which may induce modifications as glycation or methylation, which imply the addition of a small molecule to the protein, or may alterate and prevent expected modifications. In any of these cases, the proteomes of samples by an healthy individual and an affected individual should be discernible, being their mass profile altered. Therefore, among the thousands of proteins and peptides present in a serum sample, which represent its proteome, few key signals may be significant markers of the pathological state, and their search within the proteome represents a still open field of research.

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 statistical and artificial intelligence based technologies could be used for mining these data. In [Petricoin III et al., 2002], for example, genetic algorithms and self organizing 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 [Ressom et al., 2005] 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 [Lilien et al., 2003]. Several statistical methods have been compared in [Wu et al., 2003]. Nearest centroid classification has been used in several applications; in [Wu et al., 2003], for example, it is used for protein mass spectrometry while ant colony optimization has been used [Ressom et al., 2007] for peak selection from MALDI-TOF spectra. Independent component analysis has been recently used [Mantini et al., 2008] for the extraction of protein signals profiles.

In this paper we first describe a method we have implemented to extract features describing the spectra. Principal Component Analysis has been then used to further reducing features dimensionality and finally a Support Vector Machine (SVM) has been applied for classification obtaining very interesting results. The experimental data analyzed were derived from a study on women affected and unaffected by ovarian cancer. The serum samples were analyzed, as described in detail in the article by Petricoin et al. ([Petricoin III et al., 2002]), by mass spectrometry techniques, in particular by SELDI-TOF.

The paper is organized has follows. In section 2 we describe data preprocessing, features extraction/reduction and classification; then, in section 3, we describe experimental results while final discussion, some conclusions and future work are the concerns of section 4.

2 Data Preparation and Classification

2.1 *Data Preprocessing*

Before the feature selection phase, there is a preprocessing step aimed at homogenization and correction of the spectra data. The spectral data produced by a single laser shot in a mass spectrometer consists of a vector of counts. Each count represents the number of ions hitting the detector during a small, fixed interval of time. A complete spectrum is acquired within tens of milliseconds, so a typical spectrum is a vector containing between 10000 and 100000 entries. In practice, most mass spectrometers produce spectra by averaging the counts over many individual laser shots. Thus, the raw data produced by running a sample through a mass spectrometer can best be thought of as a time series vector containing tens of thousands of real numbers. Unless an entry in the vector is known to represent an actual count of the number of ions, it is usually just called an intensity and is assumed to be measured in continuous arbitrary units. Peaks in a plot of the intensity as a function of time represent the proteins or peptides that are present in the sample. A typical data set arising in a clinical application of mass spectrometry contains tens or hundreds of spectra; each spectrum contains many thousands of intensity measurements representing an unknown number of protein peaks. Any attempt to make sense of this volume of data requires extensive low-level processing in order to identify the locations of peaks and to quantify their sizes accurately. Inadequate or incorrect pre-processing methods, however, can result in data sets that exhibit substantial biases and make it difficult to reach meaningful biological results. In our experiments we applied the following preprocessing steps:

- *resampling:* Gaussian kernel reconstruction of the signal in order to have a set of d-dimensional vectors with equally spaced mass/charge values;
- *baseline correction:* removes systematic artifacts, usually attributed to clusters of ionized matrix molecules hitting the detector during early portions of the experiment, or to detector overload;
- *normalization:* corrects for differences in the total amount of protein desorbed and ionized from the sample plate.

All the above steps were implemented and applied by using the MATLAB programming environment and the Bioinformatics Toolbox.

2.2 *Feature Selection*

The feature selection and description is crucial for mass spectrometry since subsequent analysis are performed only on the selected features. Several methods have been proposed which often rely on biased data sets and can reach biological conclusion difficult to be interpreted [Baggerly and *et al.*, 2004, Sorace and Zhan, 2003]. Peak detection is the traditional method for extracting features and several techniques to identify peaks among the background noise have been proposed (see for example [Tibshirani et al., 2004]). Recently model based approaches have been revaluated for the phase of feature selection of mass spectra data [Noy and Fasulo, 2007] claiming that this approach can give a better representation of the MS signals by incorporating information about peaks shapes ans isotropic distributions. The models based methods typically perform a huge number of regressions to fit signal models to spectra. Here we adopt a hybrid method which is fast just as the peak selection methods, and at the same time tries to model the average spectrum at various scales. The basic principle adopted for the selection of features relies on the scale space theory of signal analysis [Witkin et al., 1987, Lindeberg, 1994]. The main idea of a scale-space representation is to generate a one-parameter family of derived signals in which the fine-scale information is successively suppressed. The main property of a scale-space analysis is the *causality* principle, which states that each feature at a given scale must have a cause at a previous scale. This principle preserves peaks or other feature to be artificially introduced through scales and forces the analysis to be from finer scale to coarser scales.

We assume that the peaks can be (in some way) profitably used to describe the spectrum itself; when, however, two peaks are too close they should be considered as a single maximum. Therefore the multiscale analysis can help in observing the same signal at coarser scales for feature detection and signal matching purposed [Witkin et al., 1987] as the scale increases the signal becomes coarser. Here we adopt a linear scale space which is implemented through a Gaussian kernel [Lindeberg, 1994]. In particular, here the scale is just the maximum width, σ, of the Gaussian kernel used to filter the signal. As can be seen in figure 1, some peaks collapses in a single local maximum. Clearly, as σ increases, we have smother versions of our spectra. Our feature selection phase is based on the mean of the signals at the maximum chosen scale. In particular, the local maxima of the mean of the smoothed signals are considered as the locations of the considered peaks to be used as features, see figure 2. Finally, each spectrum will be described by the mean value assumed by the original spectrum in window centered in each of the selected local maxima.

As the value of sigma increases the number of extracted features decreases, for example, at a scale of 0.1 we have 156 local maxima of the mean smoothed spectrum, whereas at scale 1.0 we have just 23 components. Therefore, as a last feature extraction step we perform a principal component analysis for dimensionality reduction (see figure 3).

2.3 *Classification*

A *learning machine* is any of such functions' estimation algorithm. The quality of the machine is evaluated in terms of the mean classification error as the number of training samples goes to infinity. The machine acts as a classifier and what we want from such a classifier is that it generalizes well, that is that it has a good balance between the *capacity* and the *accuracy* of the classifications performed, the *capacity* being the possibility of learning any given set of labels and the *accuracy* being the right classification's percentage on the training data. Rather than a classifier that has the best training accuracy, we want a classifier that do not over-fits training data and which generalize well to unseen new data. There's an extensive theoretical work on probability bounds on the actual risk of misclassification of a pattern recognition learning machine and bounds on generalization

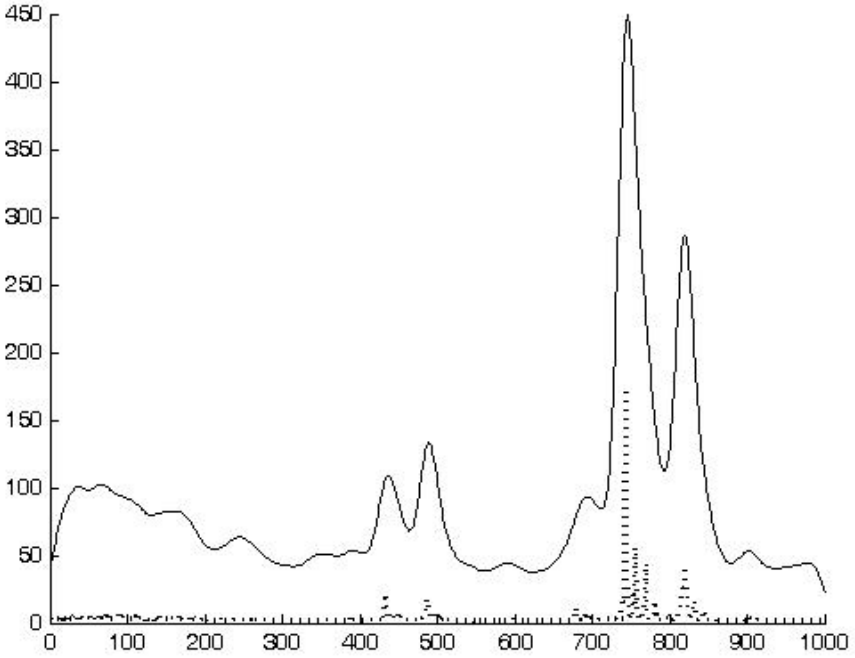

Figure 1: A spectrum (dotted) end its regularized version

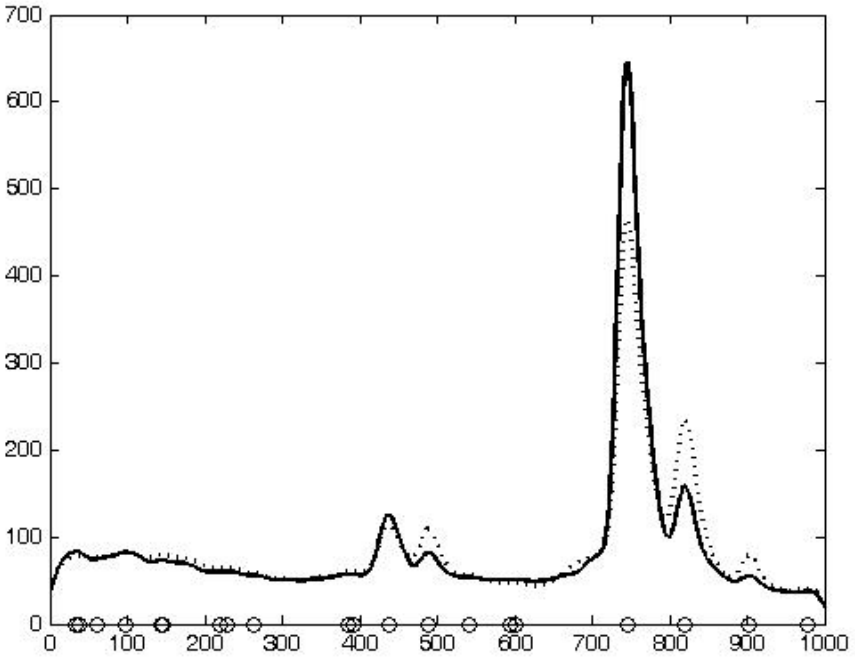

Figure 2: The curve obtained summing regularized spectra of cancer data (dotted) and the one one obtained summing the healthy data. Circles on the *m/z* axis represent local maxima points.

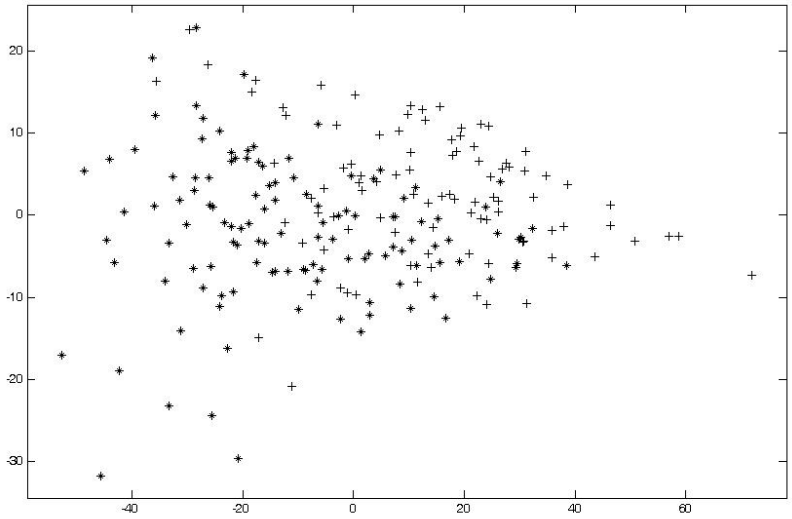

Figure 3: The data we have to classify: the first component against the second one.

performances are the subject of many institutional text books (i.e. [Vapnik, 1995]).

Support Vector Machine (SVM) is a technique proposed [Boser et al., 1992] for Pattern Recognition and Data Mining classification tasks. While at present there exists no general theory that guarantees good generalization performances of SVM, but only probability bounds on its performance accuracy, there is a growing interest in this technique due to ample literature demonstration of good performances in various heterogeneous fields [Schoelkopf et al., 1997].

The main advantage of SVM over, for example, Feed Forward Neural Networks, is that it has no local minima problems and that it has less parameters to choose. So, although there is still much study to do, i.e. in how to choose the kernel and how to extend the method to the multi-label case [Hsu and Lin, 2002], we found reasonable to test the use of this method with our data. Results reported below show very good generalization performances with our proteomic data, giving another empirical argument to its potential powerful generalization properties.

2.3.1 *Linear SVM*

In the case of linearly separable patterns on two-classes vectors it is straightforward to show the basic ideas of SVM: given a set of points in $\Re^k$ and a two-classes labels vector, SVM aims to find a linear surface that splits them in two groups according to the indicated labels, in the best possible way. Intuitively, if data are linearly separable (that is if it exists *at least* one hyperplane that splits them in two group), the problem becomes how to define and how to find *the best* possible hyperplane to do it. The SVM answer is that the best possible hyperplane is the one that maximizes the *margin*, that is the one that has maximal distance from *both* sets of points (Figure 4).

To be more rigorous, the problem of finding the hyperplane that has maximal *margin* among two sets of point differently labeled can be formulated as a constrained quadratic optimization problem

$$min(\frac{1}{2}\|w\|^2) \tag{1}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 \tag{2}$$

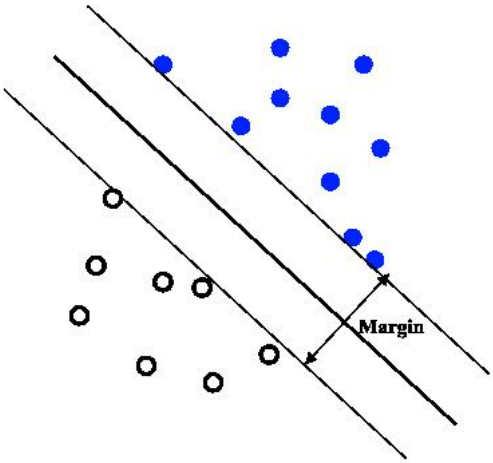

Figure 4: Maximal margin principle on a set of two classes linearly separable samples

where $y_i \in \{-1, 1\}$ are classes labels, w is the normal to the hyperplane and $2/\|w\|$is the margin, that is the distance between both sets of points.

Its dual formulation with Lagrange Multipliers is the following:

$$max(L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j) \tag{3}$$

subject to:

$$\sum_i \alpha_i y_i = 0; \;\; \alpha_i \geq 0 \tag{4}$$

where α_i are Lagrange Multipliers. This problem leads to the solution:

$$w = \sum_i \alpha_i y_i x_i \tag{5}$$

$$\sum_i \alpha_i y_i = 0; \;\; \alpha_i \geq 0 \tag{6}$$

We can note looking at the solution that the α_i are nonzero only for data lying on the marginal hyperplanes [Vapnik, 1995]. This fact has an important advantage: even removing a random subset of points not lying on the marginal hyperplanes the classification's result remains unchanged, so the technique is robust against perturbation of non-marginal points. The name of "support vectors" given to these points is due to this important property.

It can even happen that no hyperplane separates the given points. In this cases it is nonetheless possible to apply the SVM method assigning a penalty to each point in the wrong class, the weight of the penalty being a user-defined parameter. The optimization problem becomes to maximize:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{7}$$

subject to:

$$\sum_i \alpha_i y_i = 0; \;\; 0 \leq \alpha_i \leq C \tag{8}$$

where C is the "cost" of misclassified points. The only difference from the optimal hyperplane case is that the α_i now have an upper bound of C.

2.3.2 *Nonlinear SVM*

To generalize further, we could consider surfaces that are not linear and work on a different model or we could project all data non linearly in another space (possibly with infinite dimension) where they are linearly separable and perform the classification linearly in this new space, maintaining the method almost unchanged. In practice if we look at the optimization's problem formulation, we see that data appear only in the form of dot products $x_i \cdot x_j$ and hence data transformed through a function $\Phi : \Re^k \mapsto \Gamma$ (where Γ is a space of dimension $h \geq k$) appear also in the form of dot products $\Phi(x_i) \cdot \Phi(x_j)$. If we consider a dot product function $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ it is possible, in line of principle, to substitute K in formulae and to compute the solution without even knowing the form of function Φ. Such a dot product function is called *Kernel*. Whatever function that satisfies dot product's constraints can be used as *Kernel* function, and there is an active field o research in the choice of the most suitable kernel for a given problem [Cristianini and Taylor, 2004].

Another aspect of SVM is that the above discussion about *binary* classifiers can be easily extended to multiclass problems. The two most common approaches to the extension to the multilabel case are the 'one-against-one' and the 'one-against-all' [Ulrich, 1999, Hsu and Lin, 2002]. For multiclass-classification with k levels, $k > 2$, we have used the 'one-against-one' approach, in which $k(k-1)/2$ binary classifiers are trained and the appropriate class is found by a voting scheme.

3 Experimental Results

In this paper we tested support vector machines [Burges, 1998] using radial basis functions as kernel functions. In the following we will first describe the tuning of the system and then we will analyze in details the performance of our classificator.

3.1 *System Tuning*

The whole performance of our approach of course depends on (at least) four parameters: the variance σ of the Gaussian function, the size w of the window, the number n of principal components to be considered and the variance v of the RBF kernel functions; thus we decided to perform several tests having in mind to tune the whole system. Namely, we tested the classification performance having σ varying in the interval [0.1:1], the window size varying in the interval [1:21], the number of principal components used for classification varying in the interval [2:10] and v varying in the interval [0.5:5]. For each parameters' quadruplet $k-$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 of the subsets as test sets and the remaining $k-1$ subsets as training set. Each test has been repeated 10 times, so deriving that each quadruplet has been tested 100 times: in this phase, the mean correct classification value has been clearly used as quality measure.

The best mean correct classification rate (97%) has been obtained having $\sigma = 0.1$, a size of the window equal to 3, using 8 components and having $v = 3$.

3.2 *ROC Analysis*

As it is well known, a receiver operating characteristics (ROC) graph is a technique for visualizing and selecting classifiers based on their performance [Fawcett, 2006]. As it is well known, a ROC graph is a two-dimensional graph where on the X axis is plotted the *false positive rate (FPR)*

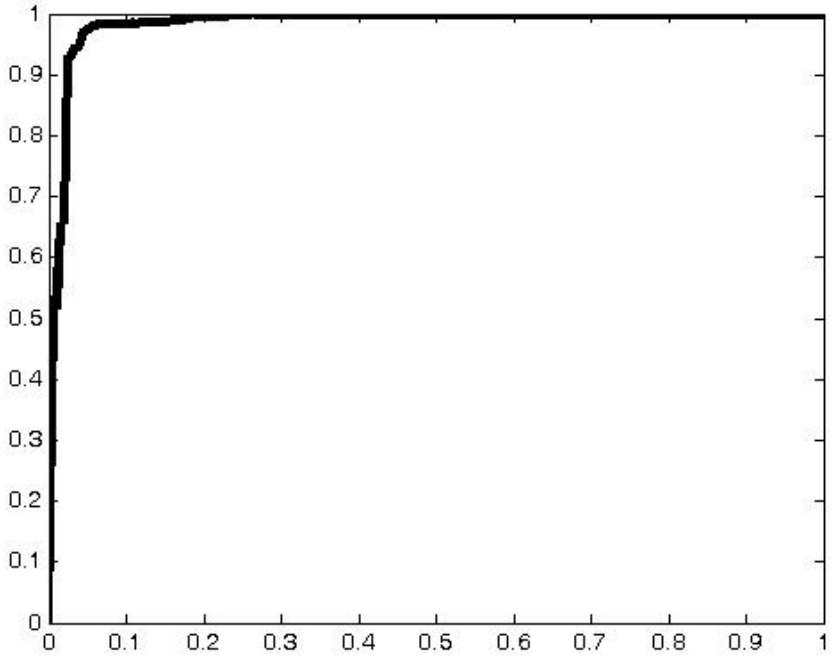

Figure 5: The vertically averaged ROC curve.

and the *true positive rate (TPR)* is plotted on the Y axis, where:

$$TPR = \frac{Positives\ correctly\ classified}{Total\ positives} \tag{9}$$

and

$$FPR = \frac{Negatives\ correctly\ classified}{Total\ negatives} \tag{10}$$

Many classifiers are designed to produce just a class decision (say true or false) and so a single confusion matrix thus deriving that these classifiers are represented by a single point in the ROC space. Some classifiers, like SVM classifiers, produce, on the other hand, a *score* that is then interpreted as the degree to which a given instance is a member of a class. Scoring classifiers are typically used with a threshold to produce a binary classifier; for each threshold we have confusion matrix and so a point in the ROC space, thus deriving a ROC curve.

According to [Fawcett, 2006], and considering the *best* system (i.e. the tuned one), we start from the 10 test sets $T_1, T_2, \ldots, T_{10}$ using 10-folds cross validation. We perfomed this set 10 times so obtaining 10 ROC graphs merged using *vertical slicing* [Fawcett, 2006]. Here each ROC graph (R_i) is treated as a function $TPR = R_i(FPR)$; the vertically averaged ROC curve is defined as the function:

$$\widehat{R}(FPR) = mean(R_i(FPR)) \tag{11}$$

Figure 5 shows the obtained graph. The area under the curve (AUC), proved to be equal the probability that the classifier ranks a randomly chosen positive instance higher than an randomly chosen negative instance, for our system is 0.986.

4 Discussion

In this paper we presented a three-steps strategy for classifying SELDI spectra. After pre-processing we described features' extraction and classification by means of SVMs obtaining very interesting results.

The features' extraction we propose is worth to be emphasized; even if in the paper it has been described without a strong theoretical formalism, the process we are using is heavily based on multi-scale analysis and the results we have obtained demonstrate the goodness of such an approach.

Several open problems need to be addressed. First of all, our method has to be tested with different data sets. Another problem we have to face is to compare our approach with other proposed solutions; this is not a trivial problem since in many cases data sets are not available and/or the software is not easy to be obtained, to be recoded or even compiled or simply used. Last, the brute force method we have used to tune the system could be easily improved using exact, heuristic or AI based techniques.

Acknowledgments

This work has been partially supported by the CNR project Bioinformatics and by Programma Italia-USA "Farmacogenomica Oncologica" Prog. No. 527/A/3A/5.

References

[Baggerly and *et al.*, 2004] Baggerly, K. and *et al.* (2004). Reproducibility of seldi-tof protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20:777–785.

[Boser et al., 1992] Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual workshop on Computational Learning Theory*.

[Burges, 1998] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 121(2):121–167.

[Cristianini and Taylor, 2004] Cristianini, N. and Taylor, J. S. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874.

[Hsu and Lin, 2002] Hsu, C. V. and Lin, C. J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13–2:415–425.

[Lilien et al., 2003] Lilien, R., Farid, H., and Donald, B. (2003). Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *Journal Computational Biology*, 10(6):925–946.

[Lindeberg, 1994] Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Kluwer Academic Publisher.

[Mantini et al., 2008] Mantini, D., Petrucci, F., Boccio, P. D., Pieragostino, D., Di Nicola, M., Lugaresi, A., Federici, G., Sacchetta, P., Di Ilio, C., and Urbani, A. (2008). Independent component analysis for the extraction of reliable protein signal profiles from maldi-tof mass spectra. *Bioinformatics*, 24(1):63–70.

[Noy and Fasulo, 2007] Noy, K. and Fasulo, D. (2007). Improved model based, platform independent feature extraction for mass spectrometry. *Bioinformatics*, 23(19):2528–2535.

[Petricoin III et al., 2002] Petricoin III, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359:572–577.

[Ressom et al., 2005] Ressom, H., Varghese, R. S., Saha, D., Orvisky, E., Goldman, L., Petricoin, E. F., Conrads, T. P., Veenstra, T. D., Abdel-Hamid, M., Loffredo, C. A., and Goldman, R. (2005). Particle swarm optimization for analysis of mass spectral serum profiles. In *GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation*, pages 431–438, New York, NY, USA. ACM.

[Ressom et al., 2007] Ressom, H. W., Varghese, R. S., K, S., Drake, Hortin, G. L., Abdel-Hamid, M., Loffredo, C. A., and Goldman, R. (2007). Peak selection from maldi-tof mass spectra using ant colony optimization. *Bioinformatics*, 23(5):619–626.

[Schoelkopf et al., 1997] Schoelkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45–11:2758–2765.

[Sorace and Zhan, 2003] Sorace, J. and Zhan, M. (2003). A data review and reassessment of ovarian cancer serum proteomics profiling. *BMC Bioinformatics*, 4:24–32.

[Tibshirani et al., 2004] Tibshirani, R., Hastiey, T., Narasimhanz, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T. (2004). Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, 20:3034–3044.

[Ulrich, 1999] Ulrich, H. G. K. (1999). *Advances in kernel methods: support vector learning*. MIT press Cambridge.

[Vapnik, 1995] Vapnik, V. (1995). *The Nature Of Statistical Learning Theory*. Springer-Verlag, New York.

[Witkin et al., 1987] Witkin, A., Terzopoulos, D., and Kass., M. (1987). Signal matching through scale space. *International Journal of Computer Vision*, pages 133–144.

[Wu et al., 2003] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19:1636–1643.