UDC 65.011

Małgorzata Plechawska
Lublin University of Technology
gosiap@cs.pollub.pl

Classification of Maldi-tof Mass Spectrometry Data in the Analysis of Cancer Patients

The article presents a case study of Maldi-Tof (Matrix-Assisted Laser Desorption Ionization – Time Of Flight) data analysis and classification. Row mass spectrometry data are preprocessed and decomposed with Gaussian Mixture Model. Gaussian mask is calculated and put at all spectra separately. In further dimension reduction RFE, PLS and T test are used. The classification is done with Support Vector Machine (SVM) method with Gaussian Radial Basis Function kernel.

Introduction

Classification is essential part of mass spectrometry data. The most common classification task is a division on two or more classes, like ill patients and healthy donors, positive or negative reaction on the medical treatment, stage of diseases. There are many papers concerning these issues on protein sequence and DNA data, microarray expressions or mass spectrometry data.

Besides the main classification very important issue is dimension reduction and features selection techniques. This task determines success of the classification because of the specificity of mass spectra data. High dimensionality of data and significantly smaller number of observations can considerably disturb classification results. Classified objects are usually represented by vectors of observed, measured or calculated features.

Supervised learning classification assumes, that there is unknown function Φ, which assign to each object of population O a label of one class. Classification process is based on the learning set U which is a subset of the hole data set O. Each element o_i of the learning set is composed of the object representation and information about its class label. This object representation is observation vector of features. The hole set is divided into c separated subsets and one subset observations are numbered among only one of c classes. Supervised learning is widely used in biomedical applications.

A construction of the prediction model

On the basic of the single learning set multiple different classifiers. The ideal situation would be to chose the proper classifier on the basic of the number of misclassifications of the new, random observation. However, in reality bad classification probabilities are unknown. They might be estimated from a validation probe. The validation probe is a random sample, independent of the learning probe, where objects' membership to classes are unknown. Misclassification probability of specific classifier is estimated with mistaken classification done by the classifier on the validation probe. Classifier evaluation should be done using observations independent of those from the learning probe. In other cases the classifier is biased.

The ultimate classifier evaluation is done with test probe. It needs to be independent of other probes and it needs to have information about objects' membership to classes. If

only one classifier is to be tested or size of the set is small, the validation probe might be omitted. In practice, the usually chosen proportion is the division: 50% on the learning probe and 25% each for the validation and test probes [1]. However, the division depends on the specificity of the data set.

The classifier makes the decision about the membership to classes on the basic of the learning probe. However in practice there is much more data among the learning set on which the classifier will work. It causes that the probability of wrong decision making is nonzero [2]. On the other hand usage of the classifier to the other data than it was built of causes that it should have ability of the generalization of learning set characteristics. In practice it means the ability of learning properties which are representative for all population with omitting those properties which are nonessential, which constituent only features of the specific learning set.

The most popular measures of classification quality are: classification accuracy (for example a proportion of correctly classified sets) and error rate (for example a proportion of misclassified sets). Important signs are also: *TP* (*True Positives*) – the number of correctly classified positive sets, *TN* (*True Negatives*) – the number of correctly classified negative sets, *FP* (*False Positives*) – the number of incorrectly classified positive sets, *FN* (*False Negatives*) – the number of incorrectly classified negative sets.

Among useful measures there are also sensitivity and specificity. The sensitivity is defined as a proportion of truly positives and false negatives results (eq. 1). It can be interpreted as the classifier ability to identify the phenomenon where it really exists.

$$sensitivity = \frac{TP}{FN + TP} \tag{1}$$

On the other hand the specificity is a proportion of truly negatives results and a sum of truly negatives and positives results (eq. 2). The specificity is interpreted as the ability to reject truly false results.

$$specificity = \frac{TN}{TN + FP} \tag{2}$$

The sensitivity and the specificity are opposed values – the increase of the one of them causes the decrease of the other one.

The significant tool characterizing classifier's features is receiver operating characteristic curve – known as a *ROC* curve. It is a chart of dependency between values: 1 – specificity and the sensitivity. Such curve is created for a specific structure of the classifier (specified type, parameters, number of input features). The total error of presented classifier remains unchanged. However, its division on values *FP* and *FN* is changed, because the *ROC* curve examines the proportion between *FP* and *FN*. In case of random division of objects the *ROC* curve will take a shape of a straight line going through the bottom left and upper right corners. The better classification results are, the more concave the curve is. The ideal situation would make the *ROC* curve go through the upper left corner of the chart.

SVM classifier

The *Support Vectors Machines* (*SVM*) is young but widely used classifier. It was proposed by V.N.Vapnik [3-5]. The idea of this method is classification with usage of appropriately designated discriminant hyperplane. Searching of such hyperplane needs Mercer theorem and optimization of quadratic objective function with linear restrictions.

If learning sub-sets are fully separable, the *SVM* idea is to find two parallel hyperplanes, which delimit the wider area do not containing any probe elements. To accept those terms the hyperplanes need to be based on some of the probe elements. Such elements are

called support vectors. The discriminant hyperplane is put in the middle of the resultant area. If learning sub-sets are not linearly separated, the penalty is introduced. The best separation is obtained for higher dimension space.

The SVM rule takes the form of (eq. 3).

$$f(x) = \operatorname{sgn}\left(\sum_{\sup.vect.} y_i \alpha_i^0 (x_i x) + b^0 \right) \tag{3}$$

where α are Lagrange's coefficients and b is a constant value. For inseparable classes the additional restrictions take the form of (eq. 4).

$$\begin{aligned} x_i w + b &\geq 1 - \xi_i, y_i = 1 \\ x_i w + b &\geq -1 + \xi_i, y_i = -1 \end{aligned} \tag{4}$$

where ξ_i is a constant value $\xi_i \geq 0$.

The more complicated classification problems are solved with use of kernel functions. Such construction enables to obtain non-linear shapes of discriminant hyperplanes. The SVM rule with kernel takes the form of (eq. 5).

$$f(x) = \operatorname{sgn}\left(\sum_{\sup.vect.} y_i \alpha_i^0 K(x_i x) + b^0 \right) \tag{5}$$

where $K(x_i x)$ is a kernel. One of the most popular kernel function is radial kernel (eq. 6).

$$K(x_i x') = \exp(-\|x - x'\|^2 / c) \tag{6}$$

Dimension reduction techniques

Input data-set for classification usually contain several hundreds or even thousands of features. From the statistical point of view using such number of features is unreasonable. There are many reduction and selection techniques available. They attempt to find the smallest data sub-set chosen with defined criteria among the hole data set. Too large number of features has an adverse impact on the classification results. Especially biological data, like mass spectrometry and microarray data fit to this characteristic. Large features number causes increase of computational complexity and lengthen of calculation time. Moreover, large number of features has an influence on low quality of classification. It is due to features correlation. This makes the analysis difficult and the diversification is hard to obtain [2]. Large number of parameters causes also large number of classifier's parameters. It increases its complexity and susceptibility on over learning and decreases its flexibility. The existence of the curse of dimensionality [6] proves, that the complexity of the classifier has an effect on the classification quality. The more complex classifier is, the higher should be the proportion between number of observation and number of features [7].

There are two types of methods:

1) features extraction – data are undergone transformation – new data set is obtained;

2) features selection – sub-set of the most optimal data is chosen.

One of commonly known features extraction methods is *Partial Least Squares* (*PLS*) [7]. It enables also classification. Features selection in *PLS* method is performed with use of both *X* and *Y* data. So it enables using structure of the hole learning data set.

The idea of *PLS* method is to find latent vectors. Using of latent vectors enables simultaneous analysis and decomposition of *X* and *Y* including covariance between *X* and *Y*. Such approach makes *PLS* a special case of *Principal Component Analysis* (*PCA*) [5].

The decomposition of X and Y is done to low-dimensional space of hidden variables. Independent variables X are decomposed according (eq. 7).

$$X = TP^T + E_x \tag{7}$$

where $T^T T = I$, I – identity matrix, T – score matrix and P – loading matrix. A product of T and P gives good estimation of X matrix.
Dependent variables Y are decomposed as(eq. 8).

$$Y = UQ^T + E_y . \tag{8}$$

The final model of PLS describing $Y \Leftrightarrow X$ regression is (eq. 9).

$$Y = X(PB_1Q^T) + E = XB + E . \tag{9}$$

SVN-RFE (*Support Vector Machine Recursive Feature Elimination*) [8] method is features selection method. Features selection is done with propagation backward method. The procedure starts with full range of input features and features are ranged successively removed. Only one feature is removed in a time. As a rang criterion *SVM* weights coefficients are used. Therefore *SVM-RFE* method is closely related to SVM classification.

In *SVM-RFE* procedure *SVM* classification might be formulated as in eq. 10.

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i^2 \\ &y_i(wz_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{10}$$

Eq. 10 is solved with eq. 11:

$$\min_{\alpha_i} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \tilde{k}(x_i x_j) - \sum_{i=1}^{n}\alpha_i \tag{11}$$

where $\tilde{k}(x_i x_j)$ is a kernel function.

The *SVM-RFE* objective function is

$$J = \frac{1}{2}\|w\|^2 \tag{12}$$

Changes in the objective function caused by features elimination may be written using the Taylor series (eq. 13). $\Delta J(i)$ in the optimal point takes the value $\Delta J(i) = (\Delta w_i)^2$, where w_i is the i^{th} removed feature.

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w^2{}_i}(\Delta w_i)^2 \tag{13}$$

Very common technique of feature selection is *T* test. The most significant features according the *T* test are chosen. For each feature a T test range is calculated with eq. 14.

$$c_i = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} \tag{14}$$

where μ_i^+, μ_i^- denote the mean values for i^{th} feature calculated for respectively positive and negative samples. Similarly σ_i^+, σ_i^- denote standard deviations and n^+, n^- denote numbers of positives and negatives learning samples.

The *T* statistics treats all feature as independent. This assumption is usually not met. However, *T* test is successfully used for protein data classification.

Characteristic of the data set

The data set presented in the paper is *Maldi-Tof* (Matrix-Assisted Laser Desorption Ionization – Time Of Flight) mass spectra data. Classification and dimension reduction methods was applied to this data set. One is serum albumin spectra obtained in the study on head cancer patients and healthy donors. The data set contains 100 data files, each of them is confirmed with four repetition. Each sample was taken from a person two times and each of those two samples was analyzed two times. The aim of the analysis is to detect peaks and find its biological interpretation. Very important part of this analysis is classification presented in this paper.

Each of the data set file contains 45 thousands of points. Typical mass spectrum is composed of two data vectors: *M/Z* value (*X* axis) and intensities (*Y* axis). Spectra, before the analysis, must be preprocessed. Preprocessing steps involve: binning, interpolation, normalization, baseline correction, normalization, denoising, peaks detection [9] and alignment [10]. One of the most important preprocessing steps is denosing, especially a baseline correction. Baseline is a special case of noise, intensifying especially in initial part of the spectrum, where *M/Z* values are low. Removal of this kind of noise flattens and averages the spectrum. The baseline correction is essential for further analysis and improves the quality of it. It is usually performed with multiple shifted windows with defined width. Normalization and interpolation are useful techniques helpful analyzing and comparing few spectra simultaneously. Interpolation is useful during the unification of measurements points [11] along with m/z axis of all spectra. Normalization [12], [13] is scaling all spectra to a single value of area under the curve. This scaling is usually done for the *total ion current* (*TIC*) value or for the constant noise. An example of analyzed spectrum with Baseline correction result is presented at Fig. 1.

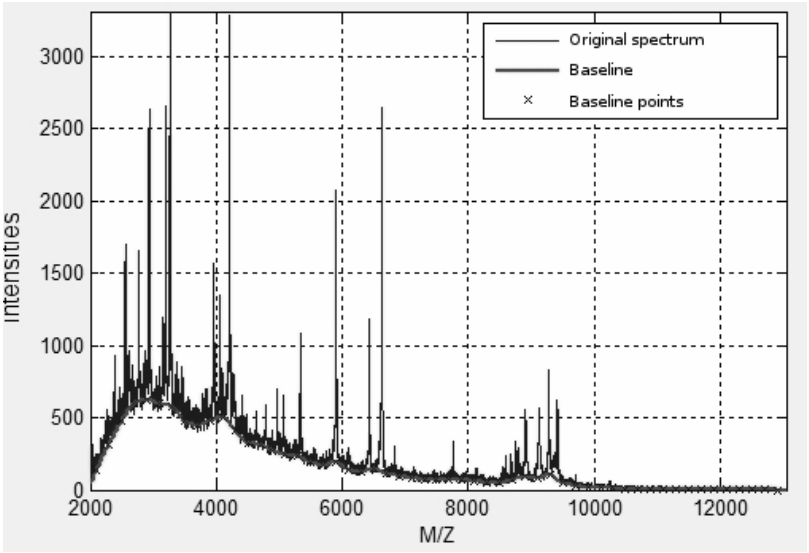

Figure 1 – An example of mass spectrum

A method which is used for mass spectra analysis is based on the Gaussian mixture decomposition. Data are modeled with *Gaussian mixture models* (*GMM*). The fitting is done with *Expectation-Maximization* algorithm (*EM*) performing maximizing the likelihood function. The analysis may be performed for several spectra simultaneously by use of the mean spectrum to make calculations faster and more efficiently.

Using of this method enables preliminary dimension reduction. After preprocessing and averaging over 4 repeated spectra of one person the mean spectrum was calculated. Next it was modeled with *GMM* with defined number of components (300).

A mixture model, as a combination of a finite number of probability distributions

$$f^{mix}(x,\alpha_1,...,\alpha_K,p_1,...,p_K)=\sum_{k=1}^{K}\alpha_k f_k(x,p_k) \tag{15}$$

where K is the number of components in the mixture and $\alpha_k, k=1,2,...,K$ are weights of particular component, $\sum_{k=1}^{K}\alpha_k=1$. Gaussian distribution (eq. 16) is given with two parameters: mean μ_k and standard deviation σ_k. Distributions in the mixture are also specified with additional parameters – weights, which determine their contribution to the hole mixture.

$$f_k(x_n,\mu_k,\sigma_k)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}\right] \tag{16}$$

The Expectation-Maximization algorithm is nonlinear method is composed of two main steps performed in the loop. The expectation step (E) consists in calculation of distribution of hidden variables (eq. 17)

$$p(k\mid x_n,p^{old})=\frac{\alpha_k^{old}f_k(x_n,p^{old})}{\sum_{\kappa=1}^{K}\alpha_k^{old}f_k(x_n,p^{old})} \tag{15}$$

The maximization step (*M*) calculates new mixture parameters values. In case of *GMM M* step is given with (eq. 18).

$$\begin{aligned}\mu_k^{new}&=\frac{\sum_{n=1}^{N}x_n p(k\mid x_n,p_{old})}{\sum_{n=1}^{N}p(k\mid x_n,p_{old})},k=1,2,...,K\\(\sigma_k^{new})^2&=\frac{\sum_{n=1}^{N}(x_n-\mu_k^{new})^2 p(k\mid x_n,p_{old})}{\sum_{n=1}^{N}p(k\mid x_n,p_{old})},k=1,2,...,K\\\alpha_k^{new}&=\frac{\sum_{n=1}^{N}p(k\mid x_n,p^{old})}{N}\end{aligned} \tag{15}$$

The decomposition results are used as a Gaussian mask which was put on every single spectrum in the data set. This gives new values consisting the spectra. Dimensions of spectrometry data decreased to the value of *GMM* components number. The result matrix obtained after those steps was: $n\times m$, where n denoted number of spectra and k – number of components.

The resultant matrix was the input data to the further dimension reduction and classification.

Results

The classification analysis was performed using all three presented dimension reduction techniques. The classification was done with *SVM* classifier with radial kernel. Tests of classification and reduction performance were done for different values of *SVM* parameters and number of selected features. To find the most accurate values, division of the data set into testing and learning subsets and classification calculations need to be repeated several hundred times. All calculations were done in Matlab environment. The *SVM* parameters are: value of box *constraints* (*C*) for the soft margin and the scaling factor (sigma). Results of multiple repetitions of *SVM* for different sigma values is presented on Fig. 2. According results the sigma value was estimated at 12 and C – as 4000.

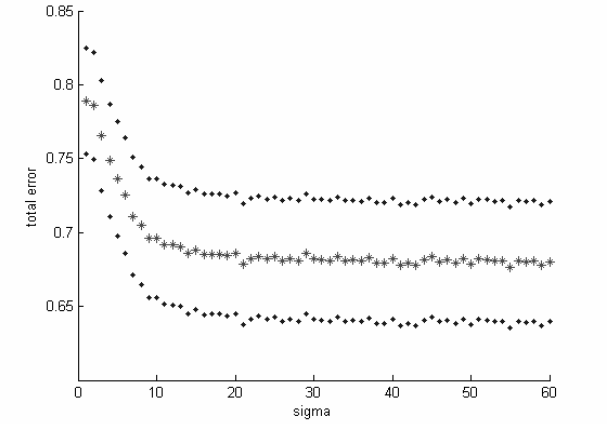

Figure 2 – Searching for optimal sigma value

If parameters are known, there is a necessity of finding optimal number of features. If there are 50-elements learning data set, number of features shouldn't be larger than 10. The results for all three types of dimension reduction techniques are presented on Fig. 3. The middle line is the obtained ratio and the upper and lower denotes the confidence interval. Similar results are obtained for *FN* and *FP* values. Fig. 4 presents typical *ROC* curve determined for the *SVM-FRE* of six features.

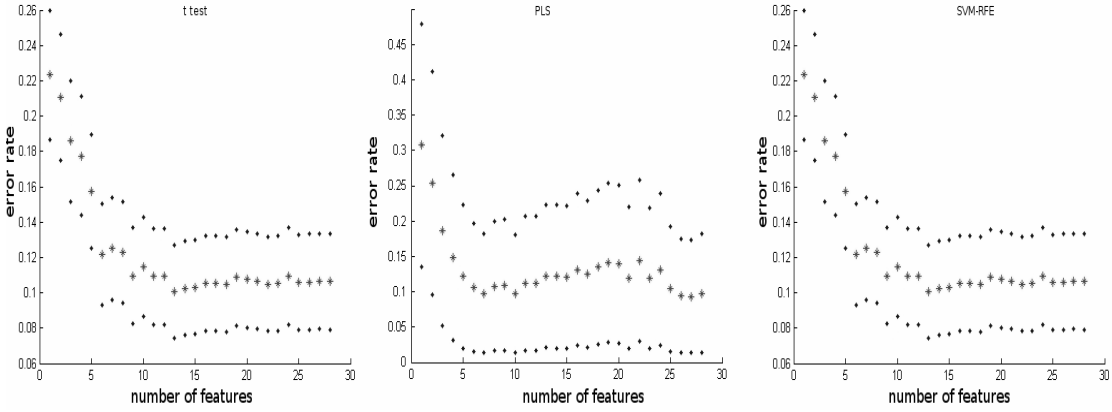

Figure 3 – Results of classification after dimension reduction

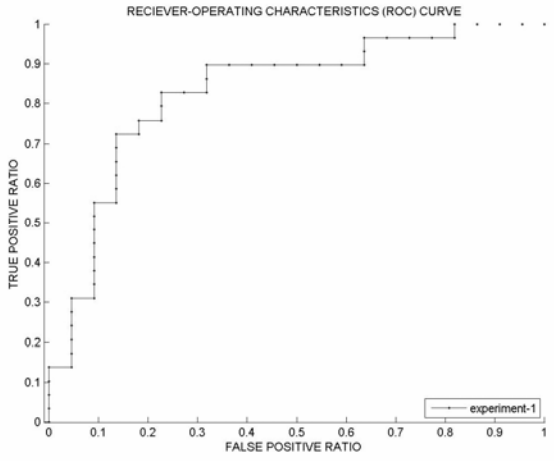

Figure 4 – The ROC curve

Conclusion

Proteomic, especially mass spectrometry data, need for special processing and analyzing. The specificity of data makes them hard to classify. Special dimension reduction techniques needs to be used. The most common technique, T test, gives good, but the weakest results. The most reliable is the *SVM-RFE* technique. It is highly connected with *SVM* classification method what makes it suitable for *MS* data. *SVM* classifier and its variants is nowadays one of the most popular classification technique among proteomic research.

Literature

1. Cwik J. Statystyczne systemy uczace sie / J. Cwik, J. Koronacki. – Warszawa : Akademicka Oficyna Wydawnicza Exit, 2008. – P. 239-245.
2. Stapor K. Automatyczna klasyfikacja obiektów / K. Stapor. – Warszawa : Akademicka Oficyna Wydawnicza Exit, 2005. – P. 35-52.
3. Vapnik V. A training algorithm for optimal margin classifiers / V. Vapnik, B. Boster, I. Guyon. – Fifth Annual Workshop on Computational Learning Theory, 1992. – P. 114-152.
4. Vapnik V.N. The Nature of Statistical Learning Theory / Vapnik V.N. – Springer, 1995.
5. Vapnik V.N. Statistical Learning Theory / Vapnik V.N. – Wiley, 1998.
6. Mao J. Statistical pattern recognition: a review / J. Mao, A.K. Jain, R.P.W. Duin. – IEEE Trans. PAMI, 2000. – Vol. 22(1). – P. 4-37.
7. Wold H. Estimation of principal components and related models by iterative least squares / H. Wold. – Multivariate Analysis. – New York : Academic Press, 1996. – P. 391-420.
8. Gene selecion for cancer classification using support vector machines / S. Barnhill, V. Vapnik, I. Guyon, et. al // Machine Learning. – Vol. 46. – P. 389-422.
9. Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum / J. Morris, K. Coombes, J. Kooman et al. // Bioinformatics. – 2005. – Vol. 21(9). – P. 1764-1775.
10. Processing MALDI mass spectra to improve mass spectral direct tissue analysis / J. Norris, D. Cornett, J. Mobley et al. // National institutes of health (USA). – 2007. – Vol. 260 (2-3). – P. 212-221.
11. TOF MS Data Graphical Preprocessing Tool / Y.V. Karpievitch, E.G. Hill, A.J. Smolka et al. // Bioinformatics. – 2007. – Vol. 23. – P. 264-265.
12. Plechawska M. Simultaneous analysis of multiple Maldi-TOF proteomic spectra using the mean spectra. SMI 2009 / M. Plechawska // Polish Journal of Environmental Studies. – 2009. – Vol.18, №. 3B. – P. 1230-1244.
13. Processing and classification of protein mass spectra / M. Hilario, A. Kalousis, C. Pellegrini et al // Mass Spectrom Rev. – 2006. – Vol. 25. – P. 409-449.

Малгожата Плехавська
Классификация Maldi-Tof масс-спектрометрических данных в исследованиях онкологических больных
Предложена классификация масс-спектрометрических данных медицинских исследований Maldi-Tof, алгоритмы и программы компьютерного моделирования, которые используются в проблемах диагностики и лечения раковых заболеваний.

Малгожата Плехавська
Класифікація Maldi-Tof мас-спектрометричних даних у дослідженнях онкологічних хворих
Запропонована класифікація мас-спектрометричних даних медичних досліджень Maldi-Tof, алгоритми та програми комп'ютерного моделювання, які використовуються в проблематиці діагностики і лікування ракових захворювань.

Статья поступила в редакцию 24.06.2010.