CONSTRUCTION AND VERIFICATION OF MATHEMATICAL MODEL OF MASS SPECTROMETRY DATA

Małgorzata Plechawska-Wójcik
Politechnika Lubelska, Wydział Elektrotechniki i Informatyki, Instytut Informatyki

Abstract. *The article presents issues concerning construction, adjustment and implementation of mass spectrometry mathematical model based on Gaussians and Mixture Models and the mean spectrum. This task is essential to the analysis and it needs specification of many parameters of the model.*

Keywords: Maldi-Tof mass spectrometry, Gaussians, Gaussian Mixture Models, SVM-RFE classification

KONSTRUKCJA I WERYFIKACJA MATEMATYCZNEGO MODELU DANYCH WIDM MASOWYCH

Streszczenie. *Artykuł przedstawia kwestie związane z konstrukcją, dopasowaniem i implementacją modelu matematycznego widm masowych opartego o rozkłady normalne i mieszaniny rozkładów oraz o widmo średnie. To zadanie jest kluczowe dla analizy, wymaga też określenia wielu parametrów modelu.*

Słowa kluczowe: spektrometria masowa Maldi-Tof, rozkłady Gaussa, mieszaniny rozkładów Gaussa, klasyfikacja SVM-RFE

Introduction

Mass spectrometry is popular, widely used technique of determination of complex data composition. It is essential technique used in proteomic, applied to identification of proteins and their dependencies in biological tissues [9, 10]. Proteomic approaches to interpretation of biological phenomenon on the level of proteins constitutes opportunities for development and advancement of medical diagnosis in many diseases area. In particular the proteomic analysis offers great promise to understanding the process of tumor development in human organism and its response to the therapy. Proteomics gives hope to the oncology development through better understanding, improvement diagnosis and development of new more efficient drugs.

The most important proteomic branches are: identification of proteins in the analysed sample, proteins features characteristic, specification of the proteins number in the sample and comparison of the proteins features.

Proteomic methods [22] must be able to deal with huge amount of data and information. High-bandwidth mass spectrometry, identification of protein complexes, studies of protein mixtures and evaluation of proteins expression requires advanced techniques. The development of proteomic techniques and tools gives opportunities of detection and analysis of proteins primarily in terms of medical diagnosis and new methods of treatments. Also statistical analysis, especially classification plays an important role in the process of the analysis of large data sets. There are still ongoing search for efficient methods of classification fulfilling the high requirements for high sensitivity and specificity. Moreover, the identification of proteins is a hard task which needs to consider many factors [34] (not always unique amino acid sequences, possible several different coding genes, existence of polymorphism genes resulting in different proteins variants, posttranslational modifications, etc). Today proteomics examines the complex protein interactions and modifications in terms of simultaneous analysis of thousands of data. Proteins identification is usually carried out by comparing the measured properties of the protein to those known and documented, available in biological (proteomic) data bases. The mass of protein derived in the mass spectrometric studies is one of the most commonly used properties. It enables low or non-invasive study of protein profiles in blood, plasma or urine.

Mass spectrometry is an analytical technique that allows accurate measurement of mass to charge ratio of the proteins. Is used to identify chemical compounds and to determine their structure and elemental composition. In proteomics, this method is used primarily to determine the composition of complex mixtures, in particular the identification of proteins. The spectrometer work consists of three stages: ionization of molecules, the selection of charged particles and their detection.

One of the most popular mass spectrometry used in the proteomic research is Maldi-Tof. It is based on using of matrix absorbing the laser beam. Sputtered and spared in the electric field ions hit the detector which determines the mass of ions on the basic of their velocities and time of flight through the spectrometer [26].

Result of spectrometric studies is presented in the form of the mass spectrum. The spectrum presents the dependence of mass-to-charge ratio (M/Z) and intensity. Intensity determines the number of ions that hit the detector in a small, fixed time interval. This interval is determined is the time resolution of the instrument and usually varies in the range 1 - 4 nanoseconds. Examples of mass spectra are presented in Fig 1.

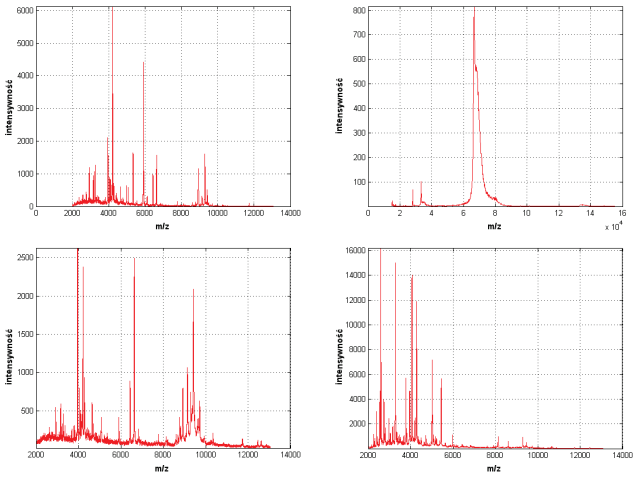

Fig. 1. Examples of mass spectra

1. Models of mass spectra analysis

A model is mathematical data representation used to present a process or phenomenon in a simplified manner. This way of the process description allows better understanding of its characteristics. For example, the model can be created through the construction of the classifier using a specific learning set and data set.

In the case of high-bandwidth mass spectrometry data important issue is to determine the purpose and tasks of the analysis. The next step is to select appropriate methods and tools for the data exploration. The analysis is performed to answer questions about patterns and the most important features in the test dataset. The phases of this analysis are presented in Fig 2.

The analysis of data groups may be based on different kind of method such as clusterization (unsupervised learning method), classification and regression. Proteomic, data are often analysed

with support vector machines (SVM [4, 35, 36]). Highly multidimensional data need also dimension reduction techniques like PCA (Principal Component Analysis) [35], PLS (Partial Least Squares) [38], ICA (Independent Component Analysis [11, 19] or MDS (multidimensional scaling) [34]. They enable to choose the most significant features from the classifier point of view. However, most of dimension reduction methods have limitations, like independence of the analyzed variables, or necessity of the normal distribution.

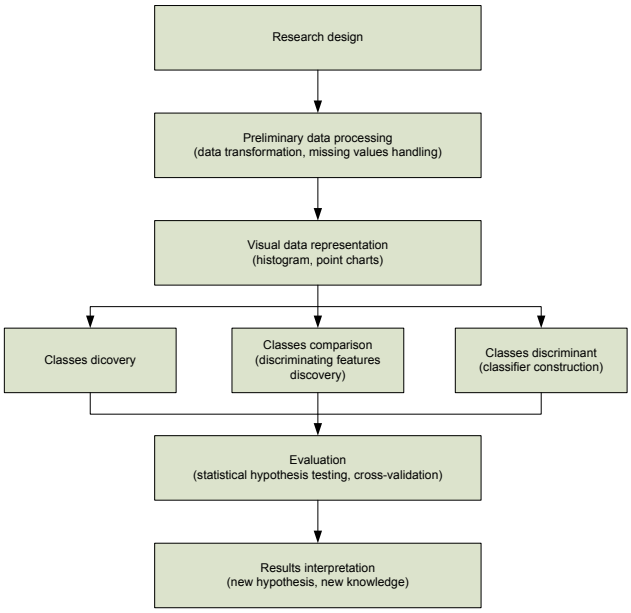

Fig. 2. Typical for proteomics construction and the data model analysis scheme

Before mass spectrometry classification analysis data need to be pre-processed and prepared. It is of great importance for the process of the further analysis and quality of obtained results.

Pre-processing steps may vary depending on the specific type of data and proposed exploration method. In the case of proteomic data coming from mass spectrometry studies noise correction, baseline correction, normalisation, spectra alignment are usually required [26, 27]. Sometimes also missing data are handled. Noise removal and baseline correction is usually necessary due to matrix noise, other sample contamination and instability. Amount of noise in the spectrum is a determinant of the spectrum quality. Noise and baseline corrections can be performed with multiple shifted windows with defined width, discrete-wavelet transformation (especially undecimated discrete-wavelet transformation, UDWT [16, 20, 21]), the least-squares digital polynomial filter (Savitzky and Golay filters) or nonparametric smoothing (locally-weighted linear regression with specified window size and type of kernel). There are also methods considering peaks shape and localization, based on the rule the higher the weight, the wider and lower peaks. The important issue is also resolution of the device. Resolution is determined the percentage of the weight. While the weight increases, the device resolution worsens. That is why spectra of low molecular mass have better resolution and their peaks are more separable. The shape of the peaks, however, varies along with the spectrum. The higher the M / Z is, the lower and wider peaks may become. The most popular method of normalization consists in scaling all spectra to total ion current (TIC) value or to constant noise. Sometimes also trimming, binning and the mean spectrum calculation is performed.

The primary source of information about proteins, their sequences and the genes encoding them constitutes biological databases. Data contained in such databases come from research experiments and their interpretation, publications and other databases. Research centres which undertake the construction and maintenance of biological databases often cooperate and exchange data. The problem of biological databases are huge growth of information and the associated need of standardization and structuring of stored information. Biological data, in particular protein data, are difficult to manage because of the continuous flow of information and lack of uniform standards and methods of naming classification. Continuous exchange of data contained in different databases enable frequent updates. On the other hand, there is a need of continuous control of data quality and consistency. There are conducted works on automation of this process and developed of new comparison and integration methods. The best-known biological databases are: UniProt, NBCI, KEGG, EXPASY, HPRD, EPO-KB.

However, before the classification and proteins searching one need to find peaks of the spectrum. There are different models used to perform this task. The most popular one is based on local maxima and minima chosen from the mass spectrum [26, 33, 39]. The real peaks can be chosen only among local maxima which are higher than a the signal to noise ratio (S/N) [15, 23, 24]. It enables detection of peaks with the highest values of intensity [1, 37]. Other methods try to distinguish between noise and the real peaks considering the shape of peaks [17, 18].

A model proposed by K. Coombes, K. Baggerly, J. Morris et al [10, 26] defines distribution of the mass spectrum of the signal in accordance with eq. 1.

$$f(t) = B(t) + N \cdot S(t) + e(t) \quad (1)$$

where f(t) is the observed signal, B(t) – Baseline, N – normalization ratio, e(t) – noise. Correct signal is defined as S(t). It can be modeled as a sum of independent, sometimes overlapped peaks, each of which corresponds to a single protein. Peak shapes can be estimated empirically on the based on physical simulation of the ToF analyzer process. The noise e(t) is defined as white noise [14]. Baggerly et al. [2] considering the inclusion of the additional noise model time-dependent factor. The method assumes using of the mean spectrum, undecimated discrete wavelet transformation (UDWT) denoising and local minima and maxima analysis.

There is also a group of methods modeling spectra with a set of member functions. Decomposition may be based on the wavelet transformations [13, 30] or composition of Levy processes [8].

Mixture models are a good way of large data sets modeling. They are usually used to model natural phenomena and biological processes. They can be also applied in image processing and clustering. Mixture models are often complex, they consist of many individual probability distributions. Mixture models allow interpret the whole population as a composite of an adequate number of sub-populations, which enable to perform detailed analysis and obtain better estimation. In practice, the most commonly used mixture models are based on Gaussian distributions. Such mixtures are known as Gaussian Mixture Models (GMM).

The main task associated with mixture models is to determine their parameters. The number of unknown parameters is 3k-1, where k is the number of Gaussian mixture's components. For each mixture component one need to estimate both its Gaussian parameters and its weight. The parameters estimation task may occur to be complex. The more components are included in the mixture, the harder and more time consuming is the estimation task.

The task of mixture models parameters solving can be treated as a missing data problem. It can be formulated as a task of determining the membership of a group of data points to one of the distributions in the mixture. This membership is unknown and must be estimated. Parameters of the model should therefore be chosen so that the data points were represented by their membership to the individual components of the mixture.

2. Expectation-Maximization Algorithm

The parameters of the mixture model need be estimated with a method, which is able to handle the missing values. In case of complex problems where the number of parameters to estimate is large, typical estimation methods, like the maximum likelihood,

are not appropriate to solve this tasks. An additional difficulty is the existence of many local likelihood function extremes. Therefore, likelihood maximisation of the data fit to the Gaussian mixture model can be performed with Expectation-Maximization algorithm (EM) [12]. The EM method assumes the existence of hidden variables. In the case of mixture models the hidden variables represent variables defining affiliation of the each observation to one of the Gaussian components.

EM algorithm is an iterative method, consisting of a repeated measures (E and M). Diagram of its operation is shown in Figure 3. E and M steps are executed in a loop until the accuracy is reached. In each subsequent step new parameters values are calculated. If the number of components is known, the initiation step of the algorithm is to determine the initial conditions.

The use of GMM for spectra modeling is based on the assumption that one peak is represented by a single Gaussian distribution. Each Gaussian defines and models one component of the spectrum. To calculate the model parameters one need to use a modified version of the algorithm Expectation-Maximization. This version is adjusted to the nature of spectrometry data.

Application of the EM algorithm requires adjustment of input data, because the standard version of EM requires a one-dimensional input vector. For the purposes of spectrometric data a weighted version of the algorithm was implemented. It considers the intensity of the spectrum characterizing the repetition number of the corresponding M / Z values.

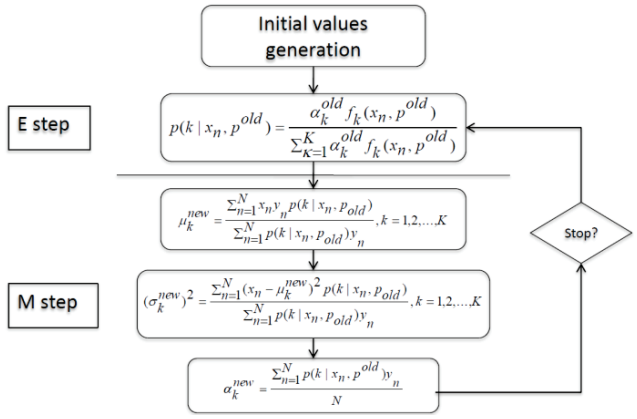

Fig. 3. The flow diagram of the EM algorithm

3. Mass spectra modeling using Gaussians and GMM

The protein mass spectra decomposition methodology might be based on Gaussian Mixture Models. This way of mass spectra processing is possible and gives good results, which has been confirmed by the results of the analysis of real data (presented later in this paper).

This mass spectral analysis method is based on peaks modelling with Gaussian distributions. Such modelling involves the appropriate choice of distributions' parameters. It enable to represent peak shapes and to model the spectra in the easy way.

Using of Gaussian distributions to model peaks allows to consider measurement errors which also can be modelled using Gaussian distributions. In most of the known methods these errors must be corrected by the difficult process of alignment of the peaks in different spectra collections. Gaussian distributions modelling does not need the aligning process, because when probability distributions are used, measurement ambiguities are allowed for a single spectrum.

A single Gaussian distributions may be applied, however, only in the case of spectra with a perfectly separable peaks. In practice, the spectra analysis is done using mixtures of distributions instead of single distributions.

Moreover, the possibility of modelling mass spectra using mixtures justifies the idea of spectra modelling with single Gaussian distributions. GMM spectra modelling is based on the assumption that the individual peaks are modelled by Gaussian distributions. However, the use of the mixture instead of single Gaussians considers the additional interaction between closely located peaks. Mass spectra reflect the number of processes occurring in the human body. These processes are often correlated and it has its reflection in a spectrum and in the lack ofseparability between the individual peaks. This fact should be upheld, as it brings information together. Application of mixtures enables considering the dependences between peaks and modelling of overlapped measurement errors present in adjacent regions of the spectrum. Separate peak identification could lead to incorrect variance estimation of individual Gaussians, since it is impossible to make their total separation. Simultaneous peak modelling can solve this problem. The measurement inaccuracy correction conducted in this manner allows to increase the quality of the assessment model. Solving the parameters of the mixture, however, needs to define many properties, such as the number of model components, the initial parameters values, the convergence criterion, calculations accuracy or the use of the mean spectrum.

4. Problems of the model implementation

Expectation-Maximization algorithm is efficient method of GMM estimation. However, to obtain repeatable, reliable results one needs to appropriately chose all parameters, such as the model parameters, its correctness, number of the model parameters, initial values, stop criterions, calculation accuracy and quality assurance.

Parameters of the pre-processing can be adjusted to the specific data before the main procedure of analysis. The order of these operations is fixed and includes: averaging technical repetition, outliers detection, baseline correction, normalization, interpolation, calculation of the mean spectrum. This order is a standard which has been developed over the last few years of research in this field. The most important parameters of the pre-processing are: the window size of the baseline correction and using (or not) the mean spectrum.

The important element of the analysis is appropriate choice of the mixture options, in particular the number of components. There is a possibility to use different methods of number of components estimating such as the BIC criterion, the basic functions of peak detection, statistical methods determining the density distribution of the parameters.

What is more, the characteristic of the EM algorithm is of great importance for the analysis. The level of the estimation task depends on the number of components and the sample size.

The EM algorithm generates some specific types of errors. The most frequent one is merging of distributions. This phenomenon occurs when at least two distributions with similar means value occurs. The merging probability is greater, if also standard deviations are similar. When the number of components is fixed, this join results in generation of additional distributions, which usually have small weight. Another type of error is generation of distributions with large standard deviation and low weight. In practice it results in long, flat distributions. Sometimes additional distributions with small standard deviation are also formed.

Another important feature of the EM algorithm is that the better and quicker adjustment are usually obtained for the components with the larger weights. This is a desirable feature, when the goal of the analysis is to find main elements of the modelled process. But it can be problematic in the case of high requirements concerning parameters fitting.

EM algorithm is an iterative, non-linear algorithm. Its convergence is fast only in the initial phase of operation (Figure 4). After a dozen of iterations the speed of approach significantly decreases and the results of successive iterations differ very little. This feature of the algorithm might longer the duration of the whole method, especially in the case of a high accuracy specified. The algorithm has also high computational complexity, especially the M step.

Despite the relatively slow convergence, premature interruption of the algorithm can cause errors. The most common problem is to find a local maximum. This situation illustrates the Figure 4. The method cannot find the parameters fulfilling the criterion of the maximum likelihood method and it remains at the point where the estimation satisfies the conditions of insufficient local maximum.

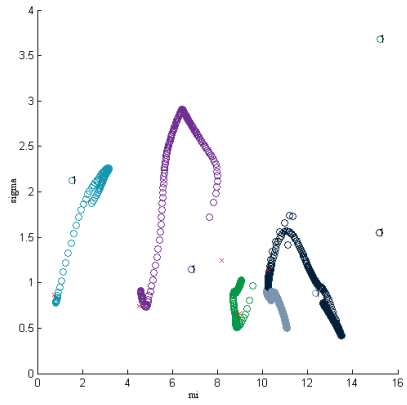

Fig. 4. The example of convergence of the EM algorithm

The local maximum problem is usually caused by improper initial values of the algorithm. To improve the method results the multiple repetitions technique can be used. It involves multiple runs of the algorithm with the initial values changed. This solution enables to choose the estimation with the maximum value of the likelihood function. This method gives good results and it effectively improves the estimation level. However, it causes extending of the duration time and higher calculation complexity.

The problem of the local maximum might be improved with careful selection of the input parameters.

The most obvious and quickest way to obtain initial parameters is the random selection. However, such a choice can be made in several ways. One is the random distribution of data into specified number of elements (this number is usually equal to the number of mixture components) [25]. The is also possibility to use one of the clustering algorithms, such as k-means or hierarchical clustering. However, using this methods might be costly, especially when it has to cover a large group of data. That is why there solutions which are based on a randomly selected samples. An alternative method of the input parameters determination is generation based on the primary peak detection method based on local extremes. Obtained results, with Gaussian arousals added, allows quick setting of the input parameters localized around the relevant procedures results.

The task of mixture's parameters estimation requires known number of components. Obtaining this number of components is important and hard task, especially in case of complicated mixtures with overlapped peaks.

One of the most efficient methods of the components number are information criteria.

Figure 5 shows a comparison of selected criteria for estimation of the components number of an example model. All criteria are based on the value of the likelihood function. They are: BIC (Bayesian Information Criterion) [32], AIC (Akaike's Information Criterion) [1], ICOMP (Information Complexity Criterion) [5, 6], NEC (Normalized Entropy Ctriterion) [7], AWE (Approximate Weight of Evidence) [3], AMIR and MIR (Minimum Information Ratio) [37].

The most stable of the analyzed criteria occurred criteria: BIC, AIC, AWE and NEC. The results presented on the graph are very similar, because the important factor in their formula is the value of likelihood function. It makes those criteria to be monotonic. Too small values of components result in significantly increase criteria values. The point there the graph is stabilizing should be treated as the optimal value. Using of the likelihood function makes the criteria value continuously growing, it is slight increase. Implementation of AWE and NEC criteria enforce maximization of the AWE criterion and minimization of the NEC criterion. AMIR and ICOMP criteria do not show so clear upward trend with an increase of the components number.

The last important issue is identify the type of convergence criterion and its accuracy [29]. The most commonly used criterion is are based on the maximum likelihood method. There are, however, other criteria like the difference between two successive values of likelihood function or different distance metrics between two successive values of the likelihood function or between successive estimates of parameters. According our simulations all tested criteria give similar results of parameters values. However, some of them run faster and require a smaller number of iterations [30]. The best results were obtained for the chi2 distance between the two constituent values of the likelihood function. Chi2 distance calculated for constituent values of parameters also gave good results in a relatively small number of steps compared. The worst results gave the distance between successive values of the parameters based on the sums of gradients between successive values of the parameters.

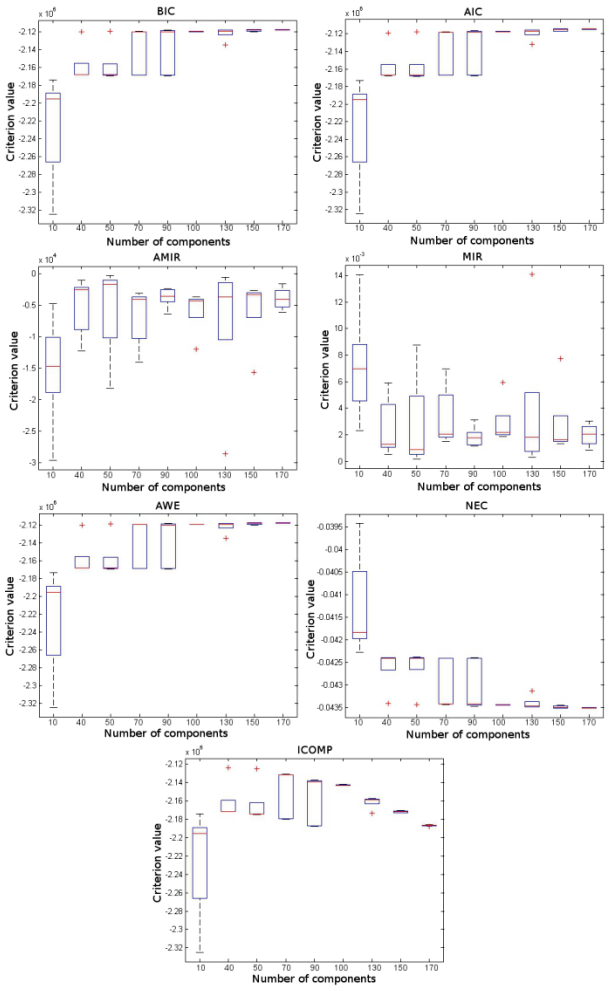

Fig. 5. Comparison of criterions of components number estimation

According our simulations all tested criteria give similar results of parameters values. However, some of them run faster and require a smaller number of iterations [30]. The best results were obtained for the chi2 distance between the two constituent values of the likelihood function. Chi2 distance calculated for constituent values of parameters also gave good results in a relatively small number of steps compared. The worst results gave the distance between successive values of the parameters based on the sums of gradients between successive values of the parameters.

The important aspect is also determination of calculation accuracy. Too low accuracy can cause the local maximum problem. Too high value will lead to an excessive increase of the computation time length.

Besides the mathematical analysis, the significant aim is identification of proteins or peptides present in the sample. Implementation of classification enables initial determination of predictive model power that allows for performing the functional analysis of identified peaks. Data classification is difficult because of a strong correlation of data combined with its high dimensionality. Row spectra, composed of many thousands of features and dozens of objects require a two-stage dimensionality reduction. The first stage is decomposition with Gaussian mask put on all spectra. This operation enable to reduce the dimensionality to several hundred of features. The second step was the selection of the most informative features in the process of dimension reduction performed with such methods like T-test, SVM-RFE or PLS. Due to the high level of correlation it is necessary to the implement classification based on features. Prepared in this way data set can be subjected to SVM learning and classifying. Appropriate adjustment of the classifier and features choice enable efficient peaks discriminatory. Selection of classification parameters and the proper number of features can be performed with Multiple Random Validation. Results performed on a medical dataset have shown that classification conducted in this way is possible and effective. Total error obtained at the level of 18% without the information leak confirms the effectiveness of the classification method.

GMM based spectra modelling, preprocessing and classification could be supplemented with biological interpretation based on dedicated databases. Integration of identification process with biological databases can be used to verify the results from the biological point of view. This verification is essential because of I type error (False Discovery Rate), typical classification task error. It is important to minimize this error, but it is impossible to do it using classification methods only. Therefore, the opportunity to conduct biological verification gives another point of view. The analysis helps to improve explaining of the processes occurring in living organisms.

5. Results

The method was tested with the set containing mice subjected to irradiation. This data set was selected due to the specificity of data. The relevant aspects are the large number of biological and technical repetitions and the possibility of treating samples as if they came from a single organism. The analyzed data set contains twelve repetitions: six biological and two technical repeats performed on five mice from one litter. The analysis conducted on this set is primarily a comparison of results obtained using two methods: with the mean spectrum and without it.

The analysis with the mean spectrum involves standard, described earlier steps:

- baseline correction – the operation is necessary due to typical character of the spectrum,
- interpolation – standardisation of points on the independent axis,
- normalization – reduction of all spectra to one common area under the curve,
- the mean spectrum calculation.

The second method used in the analysis does not require the calculation of the spectrum average. In this case the mandatory pre-processing procedure consists of only baseline correction. Normalization and interpolation, however, were also perform to standardise the data.

The graphs presented in Figure 6 illustrates the results of the decomposition analysis. Illustrative. Figure 7 shows the placement of individual spectral peaks in the confidence intervals designated for spectrum average. The graph illustrates that the medium of individual peaks in the spectra belong to the respective confidence intervals. This analysis was conducted to test the method and check its repeatability.

Figure 8 shows the box graph describing errors (differences) derived from multiple runs of a decomposition procedure for the same data set for methods based on the mean spectrum. Distances between the obtained M / Z values were considered.

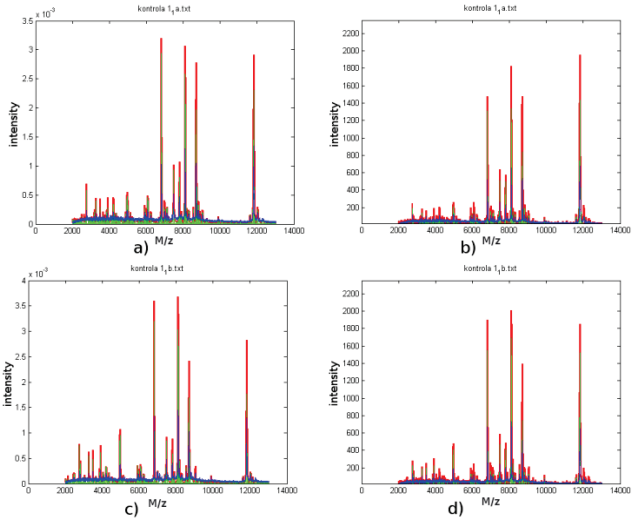

Fig. 6. Decomposition using the mean spectrum (b,d) or without it (a,c)

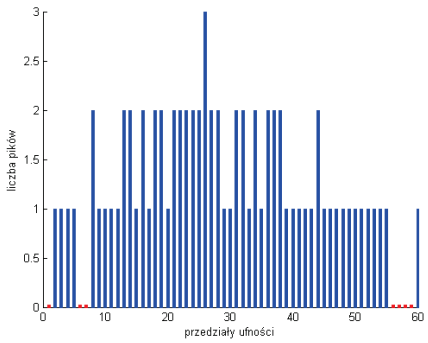

Fig. 7. Confidence intervals analysis

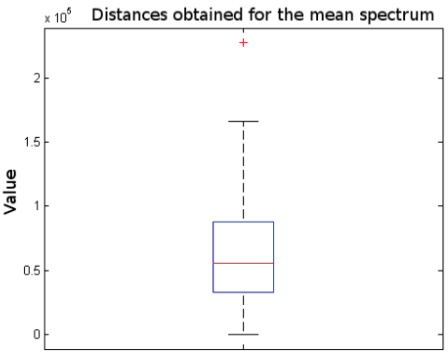

Fig. 8. Distance analysis for the method using the mean spectrum

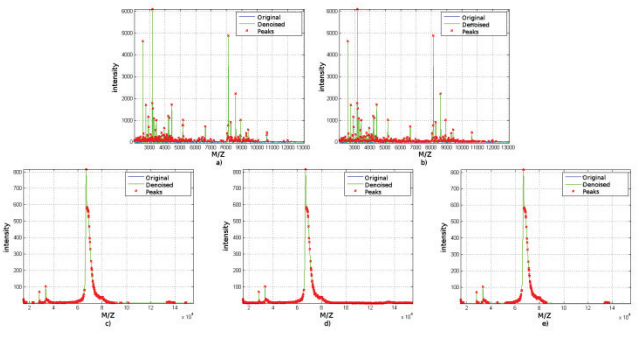

Fig. 9. Using of an alternative peaks detection method

6. Summary

Comparative analysis of methods using the mean spectrum and without it indicates that the use of the mean spectrum allows for increase efficiency and speeding up the analysis. However, results obtained for both types of analysis are similar. It indicates, that the method is reliable. Analysis also proved a high degree of reproducibility. Results of the decomposition analysis are comparable to the results obtained with other methods [28]. Moreover, in case of low level of separability it proved to be more flexible than other method. It also handle the complexity of the spectra.

Presented mass spectra processing method can be used not only for the spectra, whose peaks are narrow and do not overlap. The method also allows modeling of spectrum with a more complex structure with overlapped peaks characterized by a large variance. For these spectra, the methods based on local maxima and minima may fail. This problem is exemplified in the Figure 9. It presents using of mspeaks method (available in Matlab Bioinformatic Toolbox).

This method is based on the local maxima and minima as well as the height of the potential peaks. Those values are used to find the center point of individual peaks. Fig 9a,b present the result of the method for spectra with narrow peaks. Identification of such peaks runs smoothly and well-chosen parameters enable fast and efficient identification of peaks. Fig 9c,d,e shows the use of various configuration options for the problem of the broader spectrum with overlapping peaks. It can be seen that results are characterized with considerable redundancy of the number of identified peaks.

Bibliography

[1] Akaike H.: *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 9 s.716–723, 1974.

[2] Baggerly K.A., Morris J., Wang J., Gold D., Xiao L.C., Coombes K.R.: *A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization time of flight proteomics spectra from serum samples.* Proteomics, s. 1667–1672, 2003.

[3] Banfield J., Raftery A.: *Model-based Gaussian and non-Gaussian clustering.* Biometrics, 49 s. 803–821, 1993.

[4] Boster B., Guyon I., Vapnik V.: *A training algorithm for optimal margin classifiers.* Fifth Annual Workshop on Computational Learning Theory, s. 114– 152, 1992.

[5] Bozdogan H.: *Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher informational matrix.* Springer-Verlag,Heidelberg, 19 s. 40–54, 1993.

[6] Bozdogan H.: *On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models.* Communications in Statictics, Theory and Methods, 19 s. 221–278, 1990.

[7] Celeux G., Soromenho G.: *An entropy criterion for assessing the number of clusters in a mixture model.* Classification Journal, 13, s. 195–212, 1996.

[8] Clyde M.A., House L.L., Wolpert R.L.: *Nonparametric models for proteomic peak identification and quatification.* ISDS Discussion Paper, s. 2006–2007, 2006.

[9] Coombes K., Baggerly K., Morris J.: *Pre-processing mass spectrometry data, Fundamentals of Data Mining in Genomics and Proteomics*, W Dubitzky, M Granzow, and D Berrar, eds. Kluwer, s. 79-99. 2007, Boston.

[10] Coombes K.R., Koomen J.M., Baggerly K.A., Morris J., Kobayashi R.: *Understanding the characteristics of mass spectrometry data through the use of simulation*. Cancer Informatics, 1 s. 41–52, 2005.

[11] Comon P.: *Independent component analysis – a new concept?* Signal Processing, 36 s. 287–314, 1994.

[12] Dempster A.P., Laird N.M., Rubin D.B.: *Maximum likelihood from incomplete data via the EM algorithm.* J. R. Stat. Soc., 39,1 s. 1-38, 1977.

[13] Du P., Kibbe W., Lin S.: *Improved peak detection in mass spectrum by incorporating continous wavelet transform-based pattern matching.* Genome analysis, 22 s. 2059-2065, 2006.

[14] Dubitzky W., Granzow M., Berrar D.: *Fundamentals of data mining in genomics and proteomics.* Springer, Kluwer Boston, 2007.

[15] Fung E.T., Enderwick C.: *Proteinchip clinical proteomics: computational challenges and solutions.* Biotechniques, Suppl., 32 s. 34–41, 2002.

[16] Gyaourova A., Kamath C., Fodor I.K.: *Undecimated wavelet transforms for image de-noising.* Technical Report UCRL-ID-150931, Lawrence Livermore National Laboratory, Livermore, CA, 2002.

[17] Gentzel M., Kocher T., Ponnusamy S., Wilm M.: *Preprocessing of tandem mass spectrometric data to support automatic protein identyfication*. Proteomics, 3, s. 1597–1610, 2003.

[18] Gras R., Muller M., Gasteiger E., Gay S., Binz P.A., Bienvenut W., Hoogland C., Sanchez J.C., Bairoch A., Hochstrasser D.F., Appel R.D.: *Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection.* Electrophoresis, 20 s. 3535-3550, 1999.

[19] Jutten C., H´erault J.. *Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture.* Signal Processing, 24 s. 1-10, 1991.

[20] Lang M., Guo H., Odegard J.E., Burrus C.S., Well R.O.Jr.: *Nonlinear processing of a shift invariant DWT for noise reduction.* Proc. SPIE. Wavelet Applications II, 2491 s. 640-651, 1995.

[21] Lang M., Guo H., Odegard J.E., Burrus C.S., Well R.O.Jr.: *Noise reduction using an undecimated discrete wavelet transform.* IEEE Signal Processing Letters, 3 s. 10-12, 1996.

[22] Lewandowicz A., Bakun M., Imiela J., Dadlez M.: *Proteomika w uronefrologii - nowe perspektywy diagnostyki nieinwazyjnej?* Nefrologia i dializoterapia polska, 1 s. 15–21, 2009.

[23] Mantini D., Petrucci F., Pieragostino D., Del Boccio P., Di Nicola M., Di Ilio C., Federici G., Sacchetta P., Comani S., Urbani A.: *Limpic: a computational method for the separation of protein signals from noise.* BMC Bioinformatics, 8:101, 2007.

[24] Mantini D., Petrucci F., Del Boccio P., Pieragostino D., Di Nicola M., Lugaresi A., Federici G., Sacchetta P., Di Ilio C., Urbani A.: *Independent component analysis for the extraction of reliable protein signal profiles from Maldi-ToF mass spectra.* Bioinformatics, 24 s.63 – 70, 2008.

[25] McLachlan G.: *Finite mixture models.* John Wiley and Sons, 2001.

[26] Morris J., Coombes K., Kooman J., Baggerly K., Kobayashi R..: *Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum.* Bioinformatics, 21(9): 1764-1775. 2005.

[27] Norris J., Cornett D., Mobley J., Anderson M., Seeley E., Chaurand P, Caprioli R.: *Processing MALDI mass spectra to improve mass spectral direct tissue analysis.* National institutes of health. 2007, USA.

[28] Plechawska-Wójcik M.: *Comprehensive analysis of mass spectrometry data – a case study.* Foundations of Computing and Decision Sciences. Vol. 36 - No. 3-4, s. 275-292, 2011.

[29] Plechawska M.: *Comparing and similarity determining of gaussian distributions mixtures.* Polish Journal of Environmental Studies, 17, No. 3B s. 341–346, 2008.

[30] Polanska J., Plechawska M.: *Comparison of convergence criterions used in expectation-maximization algorithm.* Symbiosis, 2008.

[31] Randolph T., Mithcell B., McLerran D., Lampe P., Feng Z.: *Quantifying peptide signal in maldi-tof mass spectrometry data.* Molecular & Cellular Proteomics, 4 s. 1990–1999, 2005.

[32] Schwarz G.: *Estimating the dimension of a model.* Annals of Statistics, 6 s. 461–464, 1978.

[33] Tibshirani R., Hastiey T., Narasimhanz B., Soltys S., Shi G., Koong A., Le Q.T.: *Sample classification from protein mass spectrometry, by 'peak probability contrasts'.* Bioinformatics, 20 s. 3034 – 3044, 2004.

[34] Tversky A., Hutchinson J.W.: *Nearest neighbor analysis of psychological spaces.* Psychological review, 93(1) s. 3–22, 1993.

[35] Vapnik V.N.: *The Nature of Statistical Learning Theory*. Springer, 1995.

[36] Vapnik V.N.: *Statistical Learning Theory*. Wiley, 1998.

[37] Windham M.P. Cutler A.: *Information ratios for validating cluster analyses*. Journal of the American Statistical Association, 87 s. 1188–1192, 1993.

[38] Wold H.: *Estimation of principal components and related models by iterative least squares*. Multivariate Analysis, s. 391–420, 1966.

[39] Yasui Y., Pepe M., Thompson M.L., Adam B.L., Wright G.L., Qu Y., Potter J.D., Winget M., Thornquist M., Feng Z.: *A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection.* Biostatistics, 4 s. 449-463, 2003.

Ph.D. Eng. Małgorzata Plechawska-Wójcik
e-mail: gosiap@cs.pollub.pl

Malgorzata Plechawska-Wójcik received her PhD in computer science in 2011 from Silesian University of Technology, Gliwice.
Her research interests include bioinformatics and medical diagnosis support systems, software engineering and biological databases. Currently she performs her research projects at Silesian University of Technology, Gliwice, Poland. She is an assistant professor at Lublin University of Technology.

Artykuł recenzowany