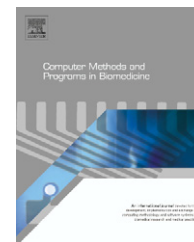




ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

An intensity-region driven multi-classifier scheme for improving the classification accuracy of proteomic MS-spectra

Panagiotis Bougioukos^{a,*}, Dimitris Glotsos^b, Dionisis Cavouras^b, Antonis Daskalakis^a, Ioannis Kalatzis^b, Spiros Kostopoulos^a, George Nikiforidis^a, Anastasios Bezerianos^a

^a Department of Medical Physics, School of Medicine, University of Patras, GR-26504 Patras, Rio, Greece

^b Medical Signal and Image Processing Lab, Department of Medical Instruments Technology, Technological Educational Institute of Athens, Greece

ARTICLE INFO

Article history:

Received 1 August 2008

Received in revised form

26 October 2009

Accepted 4 November 2009

Keywords:

Classification

Ovarian cancer

Pre-processing

ABSTRACT

In this study, a pattern recognition system is presented for improving the classification accuracy of MS-spectra by means of gathering information from different MS-spectra intensity regions using a majority vote ensemble combination. The method starts by automatically breaking down all MS-spectra into common intensity regions. Subsequently, the most informative features (m/z values), which might constitute potential significant biomarkers, are extracted from each common intensity region over all the MS-spectra and, finally, normal from ovarian cancer MS-spectra are discriminated using a multi-classifier scheme, with members the Support Vector Machine, the Probabilistic Neural Network and the k -Nearest Neighbour classifiers. Clinical material was obtained from the publicly available ovarian proteomic dataset (8-7-02). To ensure robust and reliable estimates, the proposed pattern recognition system was evaluated using an external cross-validation process. The average overall performance of the system in discriminating normal from cancer ovarian MS-spectra was 97.18% with 98.52% mean sensitivity and 94.84% mean specificity values.

© 2009 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Ovarian cancer is the fifth leading cause of cancer-related deaths in women in the United States [1]. Despite the fact that ovarian cancer is ten times less common than breast cancer, it is three times more lethal [1]. The high mortality rate of ovarian cancer might be related to the lack 1/of a screening strategy to detect early stage disease and 2/of the small number of specific symptoms of the disease in the early stages [1].

A common screening strategy followed for the ovarian cancer early stage detection, mainly in the high risk population groups [1], includes annual pelvic examinations, transvaginal ultrasound and serial measurements of the biomarker 125 (CA-125). Nevertheless, the latter biomarker, exhibits a sensitivity of less than 60% in early stages of the disease [2,3]. Thus, more sensitive biomarkers are required for accurate the detection of ovarian cancer.

Proteomics has been shown an important tool for biomarker discovery aiming towards effective and reliable diagnosis of ovarian cancer [4]. An important aspect in

Abbreviations: k -NN, k -nearest neighbor; PNN, probabilistic neural network; SVM, support vector machines; MV, majority vote.

* Corresponding author. Tel.: +30 2610 996114.

E-mail address: bougiouk@upatras.gr (P. Bougioukos).

0169-2607/\$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.

doi:10.1016/j.cmpb.2009.11.003

proteomics data analysis is the investigation of mass spectrometry (MS) spectra profiles, which concerns the selection and utilization of a number of m/z values, which encode information concerning the correlation of proteins or peptides with various diseases.

Significant research has been conducted in search for more accurate methods for screening ovarian cancer. Most of these studies have been tested on the ovarian cancer proteomic dataset (8-7-02) [5] and can be categorized into two main groups: 1/studies that have used all MS-spectrum's m/z values [4,6–13] and 2/studies that have used the MS-spectrum's m/z values which were greater than 1000 [14–16]. In the first category, research efforts have been presented using various feature selection approaches, such as Separation measures, Envelope eccentricity, entropy, Pearson correlation, and Signal to noise correlation [6], signal to noise statistics [12], entropy methods and statistical hypothesis testing [10,11]. Moreover a panel of classification methods have been employed such as genetic algorithms [8], Logical Analysis of Data [6], Nearest Neighbour classifiers [11,13], genetic algorithms coupled with cluster analysis [4], decision trees [9] and support vector machines (SVM) [7,10–12]. Results have shown classification accuracies from 92% to 100% [4,6–13,15]. In the second category (m/z values greater than 1000), authors in [15] have employed a pre-processing scheme that has comprised, baseline subtraction, average spectrum generation, spectra alignment, peak detection and potential biomarker identification attaining 94% accuracy using the Adaboost classifier, for m/z values greater than 1500. In [14] a method has been used employing combinatorics and optimization-based methodology of Logical Analysis of Data. The authors have applied the method excluding each time a part of the spectrum investigating the discriminatory efficacy of smaller and larger m/z values. For m/z values greater than 1000, results have shown 89% sensitivity and 93% specificity, for m/z values greater than 2000 accuracies were 89% sensitivity and 92% specificity, whereas for m/z values greater than 3000 predictions rates were significantly reduced. In [16] a two-sided wilcoxon test has been used for feature reduction; classification rules have been constructed using discriminant analysis. For m/z values greater than 2000, the method has resulted for a single data partition into one training and one test dataset to 100% sensitivity and 95.7% specificity.

In this study, a novel method is presented for the feature selection and classification of proteomic MS-spectra with application to ovarian cancer. The method starts by automatically breaking down all MS-spectra into common intensity regions. Subsequently, the most informative features (m/z values) are extracted from each common intensity region over all MS-spectra using a pattern recognition system. Finally, the normal from the ovarian cancer MS-spectra are discriminated using a multi-classifier scheme. The proposed method was evaluated on the public available proteomic dataset (8-7-02) [5]. Only the m/z values greater than 1500 were considered, following the findings in [17], which suggest that it is sometimes difficult to extract reliable information by analyzing the part of MS-spectra with m/z values below 1500 due to distortion of this part of MS-spectra by energy-absorbing-molecules. The proposed method differs from others in three key issues. (a) *Methodology*: The novel concept of feature extraction from

common intensity regions over all MS-spectra is introduced, (b) *Reliability*: The proposed method, in contrast to previous studies, was evaluated using an external cross-validation process. Thus, results may be considered reliable and indicative of the generalization capability of the proposed method to unseen data. Previous studies have either split the entire dataset into one training and one test set [15], or have used k-fold validation procedures to estimate the performance of the classification mechanisms employed [14,16]. Splitting data into only one training and test set cannot be considered as an optimal process for robust estimates of the classifier performance [18]. On the other hand, k-fold validation methods are subjected to feature selection bias as has been shown by Ambroise and McLachlan [18], resulting in this way to optimistically estimates of the classifiers' performance. The external cross-validation adopted in this study, takes into account and corrects for the feature selection bias, safeguarding the reliability of classification performance estimates. (c) *Accuracy*: Compared to previous studies that have analyzing the part of MS-spectra with values greater than 1000, we will show that the proposed method results to the highest classification estimates, which at the same time, may be considered as reliable of the performance of the method to unseen MS-spectra.

2. Materials and methods

2.1. Dataset

MS-spectra from the ovarian-cancer dataset (8-7-02) were obtained from the National Cancer Institute Clinical Proteomics Database [5]. The samples were processed with a robotic device. MS-spectra were produced using the WCX2 protein chip and an upgraded PBSII SELDI-TOF mass spectrometer. The dataset comprised 91 controls and 162 ovarian cancer cases.

2.2. Data partition-system evaluation design

The proposed method was evaluated in terms of the external cross-validation method which provides a nearly unbiased estimate of the prediction error [18]. Accordingly, the MS-spectra dataset was randomly separated into 2 subsets: a training dataset (70% of the data) used for generating (designing) the classification scheme, and a testing dataset (30% of the data) used for assessing its predictive performance on unseen MS-spectra. Data pre-processing, noise estimation, peak detection, peak common intensity regions generation and peak alignment were performed on the training dataset. The overall classification accuracy for the testing dataset was recorded. This process was repeated 10 times for each split of the external cross-validation method into training and testing datasets.

2.3. Baseline subtraction – normalization – smoothing

All MS-spectra exhibited a baseline drift due to chemical and electronic noise [19]. Accordingly, the baseline drift of each MS-spectrum was estimated by using multiple shifted win-

dows of 200 bins (m/z values) size. Spline approximation was used to regress the varying baseline. The regressed baseline was subtracted from the spectrum, yielding a baseline corrected spectrum [20].

Normalization was used to reduce variation in signal intensity between spectra [21]. In order to be consistent with [9,11,12,16], the spectra of the training dataset were normalized according to Eq. (1):

$$NV = \frac{(V - \text{Min})}{(\text{Max} - \text{Min})} \quad (1)$$

where NV is the normalized intensity, V is the spectrum's intensity to be normalized, Min and Max are the minimum and maximum intensities of all MS-spectra in the training dataset. These values (Min , Max) were used to normalize the spectra of the corresponding testing dataset employing.

Smoothing was employed for reducing spectrum's spikes that appear to constitute peaks which are not replicates at all spectra [21]. The smoothing process was performed using the Lowess smoothing technique [22].

2.4. Noise estimation

A local noise estimation procedure was followed for noise removal. Accordingly, a sliding, non-overlapping window scanned each spectrum across the m/z value axis calculating at each position the local histogram. Local noise was computed according to Eq. (2) for each position of the sliding window, considering only intensity values below the 90th [23] percentile of the local histogram. Following, those spectral points exhibiting values below the estimated *local.noise* Eq. (2) level were considered as noise and were omitted from further analysis. The width of the sliding window was fixed, comprising 1% of all spectral points, whereas the size of the sliding window was variable, since the distance between adjacent spectral points is not equal. Local noise estimation was performed on the training dataset (Section 2.2).

$$\text{local.noise} = \text{mean.value} + \text{std} \quad (2)$$

2.5. Peak detection

Peaks were detected using a simple differentiation method [24] performed for each denoised spectrum. Selected peaks were considered as encoding information concerning potential biomarkers and were used for further analysis.

2.6. Peak-intensity regions generation

Each MS-spectrum's selected peaks were broken down into 5 intensity regions. The borders of each region were experimentally determined for optimum classification performance defined at 0, 50, 60, 70 and 80 percentiles of each spectrum's normalized intensity histogram. Specifically, a simple algorithm was designed that determined the intensity thresholds as follows:

First, the ovarian MS-spectra were split once into one training dataset (70% of the MS-spectra randomly selected) and one test dataset (the remaining 30% of the MS-spectra). Sec-

ond, a mean spectrum was formed [25] from the training dataset. Third, "peak detection" on the mean spectrum was performed for locating existing peaks in the mean spectrum. Fourth, equidistant intensity thresholds were initially set and their values were appropriately modified so that the resulting intensity regions would each contain approximately an equal number of peaks. Finally, the so determined mean spectrum's threshold levels were rounded to the nearest ten. The number of intensity regions for optimum classification performance was determined by experimentation, by repeating the whole classification process (see Section 2.8) for 1, 3, 5, or 7 intensity regions. It was found that by splitting the ovarian MS-spectra into 5 regions provided the highest classification accuracy.

2.7. Peak alignment

A peak alignment process succeeded for each common intensity region separately, in order to counterbalance possible shifts of the same peaks between different spectra, a phenomenon appearing as a combined effect of chemical and electronic noise [19]. Accordingly, peaks that were found to differ less than 0.1% of relative mass [4] (m/z value $\pm m/z$ value $\times 0.001$) were classified as of corresponding to the same m/z value. Resulted m/z values (features) were subsequently used for classification in order to a/discriminate normal from cancer ovarian MS-spectra and b/select m/z values reflecting potential biomarkers.

2.8. Classification

To facilitate notation, from this point and on, each m/z value will be referred to as *feature*, each set of m/z values describing a single MS-spectrum will be referred to as *spectrum sample*, and each of the 5 common intensity regions will be referred to as *group*.

The classification scheme was designed as follows: At each repetition of the external cross-validation process, during which data were split into training and testing datasets (Section 2.2), the peaks of the first group of the training dataset were used as input to three classifiers, namely the Support Vector Machine (SVM) [26], the Probabilistic Neural Network (PNN) [27] and the k -Nearest Neighbour (k -NN) [27]. For each classifier, the optimal feature subset that maximized the classification accuracy was determined, employing the sequential forward selection method [27]. Among the three classifiers, the one that provided the maximum classification accuracy was finally selected to discriminate, in the testing dataset, normal from cancer spectrum samples. The same process was repeated for the remaining (four) groups.

The end result of the above process was the selection of a single classifier design (classification algorithm and features) for each group. Thus, five different classifier's designs were obtained. Subsequently, the 5 classifiers' designs were combined on a single ensemble scheme [28] using a majority vote rule (Eq. (3)), in order to discriminate normal from cancer ovarian spectrum samples:

$$MV(\mathbf{x}) = \sum_{g=1}^5 d_g(\mathbf{x}) \quad (3)$$

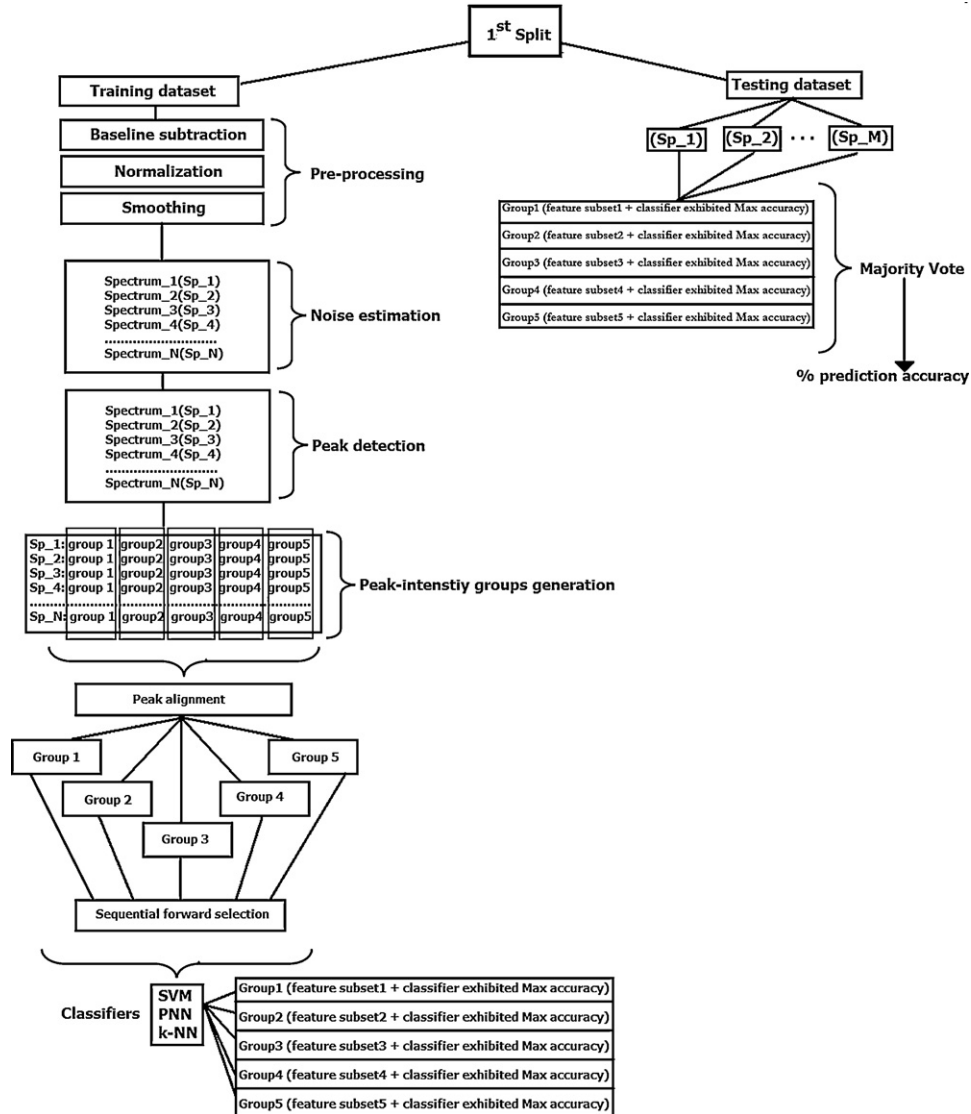


Fig. 1 – The proposed classification scheme employed at each data partition (10 splits).

where MV is the majority vote decision concerning an unknown spectrum sample \mathbf{x} , $d_g \in [-1, -1]$ is the binary decision result at the g -th peak-intensity group, where $g = 1, 2, \dots, 5$. The unknown spectrum sample is then classified as normal if $MV(\mathbf{x}) > 0$ or cancerous otherwise.

Classification results for the 10 data splits at each repetition of the external cross-validation process (see Section 2.2) were averaged by computing their mean and standard deviation values.

In the classification scheme (Fig. 1) the classifiers employed were:

1/SVM [26]; The discriminant function of the SVM classifier is given by Eq. (4):

$$D^{(SVM)}(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^N a_i y_i K(\mathbf{x}, \mathbf{x}_i) + w \right) \quad (4)$$

where \mathbf{x} is the unknown sample, \mathbf{x}_i is the i -th training feature vector, y_i is the corresponding class label $[-1, +1]$, N is the

number of samples and K is the kernel function [26]. The optimization problem of calculating parameters α_i , was solved by using the function *quadprog* provided by the Matlab™ (The Mathworks, Inc., Natick, MA) optimization toolbox. In order to select the most suitable kernel for the ovarian dataset classification four kernels, linear, second order polynomial, third order polynomial and radial basis function were tested. Best results were obtained using the linear kernel (Eq. (5)). The same observation was previously reported for the same ovarian dataset in [12].

$$K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x} \cdot \mathbf{x}_i \quad (5)$$

2/PNN [27]; The discriminant function of the PNN classifier is given by Eq. (6).

$$D_k^{(PNN)}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \sigma^d} \frac{1}{N_k} \sum_{i=1}^{N_k} \exp \left[-\frac{(\mathbf{x} - \mathbf{x}_{ki})^T (\mathbf{x} - \mathbf{x}_{ki})}{2\sigma^2} \right] \quad (6)$$

Table 1 – Classification accuracies (mean and standard deviation) for the majority vote rule scheme over the ten runs of the external cross-validation process.

	Prediction accuracies (%)		
	Overall accuracy	Specificity	Sensitivity
Mean	97.18	94.84	98.52
Standard deviation	2.61	4.08	1.91

where σ is the spread of the Gaussian activation function (experimentally determined for best performance to 0.5), N_k is the number of samples of class k , d is the dimensionality of samples, \mathbf{x}_{ki} is the i -th sample of class k , and \mathbf{x} is the unknown sample. The latter is classified to the class with the highest discriminant function value.

$3/k$ -NN [27]; The k -NN classifier computes the distances between an unknown sample and the samples of each class (neighbours) and accordingly, classifies the unknown sample to the class that most of the k neighbours belong to. As a distance metric, the Euclidean metric was used. The k parameter was experimentally determined for optimal classification performance equal to $k=17$.

Best PNN and k -NN parameters for best performance were experimentally determined as follows: Considering the first dataset-split into *training* and *test* datasets (see Section 2.6), multiple trials in individual classifier design were performed, each time using different classifier parameter values and employing the leave-one-out method for evaluating the performance of the classifier on the training dataset alone. Here, the whole dynamic intensity range of the normalized MS-spectra (i.e. one intensity region) of the training dataset was considered.

3. Results

According to the external cross-validation process, the mean overall performance of the system in discriminating normal from cancer ovarian MS-spectra was 97.18% (Table 1). The mean values of the sensitivity and specificity were 98.52% and 94.84%, respectively. These accuracies were obtained using the majority vote rule, which was composed using the most efficient classifier at each intensity region (group). Specifically, for

Table 3 – Most frequently found m/z values for all repetitions of the external cross-validation.

Group1	Group2	Group3	Group4	Group5
Common intensity regions				
1671.2718	1617.5351	2199.6676	3203.7495	10946.194
1789.2664	1657.1856	2277.3762	3247.2115	11035.715
1961.661	1670.8902	2437.3022	5523.359	11218.819
2080.5218	1962.2122	2796.5803	9857.6911	11506.3085
2307.7616	2080.9474	2985.9891	10020.041	11593.5665
2666.3614	2307.3134	3161.631	10040.6035	11628.2612
3070.4668	2543.0108	3203.7497	10061.187	12133.8002
3532.6867	2559.0366	3673.5666	10126.82	12722.0955
3642.5224	2666.3614	3745.7544	11043.559	12749.478
3993.6132	2795.5934	3822.682	11385.4985	12893.192
4186.3786	3070.6396	4264.0255	11624.737	13162.685
4596.2886	3532.4088	5034.386	12132.2555	13191.6095
5274.4582	3648.1569	5276.492	13921.7402	13320.5455
5958.8662	3673.0013	5771.2609	14152.4607	13937.162
7054.7185	4266.463	6602.8576	15711.2158	15710.6308
7250.8513	4874.982	6655.2851	17130.407	
7378.9559	5039.9044	7086.8919	18478.782	
7535.7043	5275.8139	17101.1068		
7965.7759	6039.8236			
9168.3	6812.6608			
9436.5226	6847.7542			
10525.1005	7244.8914			
	7383.7672			
	9531.0427			

the 1st group the best classifier was the SVM, for the 2nd group the SVM was found most efficient in most repetitions, and for the remaining 3 groups all classifiers were approximately equally efficient as shown in Table 2. Table 3 illustrates the most frequently found m/z values at each intensity group in the ten repetitions of the external cross-validation procedure.

Table 4 shows the classification results obtained employing the whole dynamic range of the normalized MS-spectra, i.e. without incorporating the peak-intensity regions generation step and the ensemble classification scheme. In this case, only individual classifiers could be used to assess the performance of the system, since the majority vote combination rule was designed to combine information from different intensity regions. It can be observed from Tables 1 and 4 that the performance of the system employing the majority vote combination rule was significantly higher (97.18%) as compared

Table 2 – Performance and structure of the majority vote scheme for each repetition of the external cross-validation process using the majority vote combination rule.

External cross-validation repetition	Prediction accuracies (%)			Structure of the majority vote combination rule	Common intensity regions				
	Overall	Specificity	Sensitivity		Group1	Group2	Group3	Group4	Group5
1	98.82	96.77	100	Structure of the majority vote combination rule	SVM	KNN	SVM	KNN	SVM
2	98.82	96.77	100		SVM	SVM	SVM	KNN	SVM
3	98.82	96.77	100		SVM	SVM	KNN	PNN	PNN
4	97.65	96.77	98.14		SVM	SVM	SVM	SVM	SVM
5	98.82	96.77	100		SVM	SVM	PNN	SVM	KNN
6	96.47	93.55	98.15		SVM	SVM	KNN	SVM	PNN
7	98.82	96.77	100		SVM	SVM	KNN	KNN	SVM
8	95.29	93.55	96.30		SVM	SVM	KNN	PNN	SVM
9	90.59	83.87	94.44		SVM	SVM	SVM	KNN	KNN
10	97.65	96.78	98.15		SVM	SVM	SVM	SVM	KNN

Table 4 – System's performance using individual classifiers and excluding the peak-intensity regions generation step.

External cross-validation repetition	SVM (%)	PNN (%)	KNN (%)
1	89.41	91.76	94.12
2	96.47	94.12	95.29
3	89.41	91.76	95.29
4	90.59	85.88	91.76
5	96.47	91.76	91.76
6	97.65	88.24	89.41
7	95.29	89.41	92.94
8	95.29	89.41	96.47
9	90.59	88.24	87.06
10	88.24	83.53	95.29
Mean accuracy	92.94	89.41	92.94
Standard deviation	3.60	3.14	2.99

to the performance of the best individual classifiers (92.94%, 89.41%, 92.94% for the SVM, PNN, k-NN, respectively).

4. Discussion

The goal of this study was to develop a structured pattern recognition system designed to interpret information from each intensity region for optimum separation of normal from ovarian cancer MS-spectra. The peak-intensity regions generation step proved efficient in boosting up the performance of the system under the Majority Vote combination rule. Previous studies [23,29] have utilized similar classifier combination schemes, but from a different perspective. According to these studies, the predictions of individual classifier members of the combination schemes were based on coding aggregated information extracted from each MS-spectrum. On the other hand, individual classifier members in this study were designed to code information from only a part of each MS-spectrum, the so-called intensity region, the borders of which were common for all MS-spectra. Subsequently, only one individual classifier (the one that gave the maximum classification accuracy) was selected to represent each common intensity region; the five resulting classifiers, corresponding to each one of the five common intensity regions, were combined to form the ensemble prediction rule. Results have shown that prediction rules based on common intensity regions were uncorrelated, since the proposed Majority Vote rule gave significant higher accuracy compared to individual classifiers (92.94% for both k-NN and SVM, and 89.41% for the PNN).

Another issue that has to be mentioned is that only the m/z values greater than 1500 were considered in this study to avoid possible sources of distortion of the 'lower' part of MS-spectra by energy-absorbing-molecules. The latter distortion might lead to optimistically biased estimates as it has been shown in [17]. Results presented in this study are in line with other studies that have investigated m/z values above 1000; however, our results might be regarded as more reliable. Indeed, in this study the external cross-validation method was used to estimate the performance of the prediction rules (97.18% overall accuracy, averaging ten independent splits of data into training and testing subsets), in contrast to other studies that have used validation methods, according which the training data were also used as testing data (89% specificity and 97% sensitivity for m/z values >1500 [15], 89% sensitivity and 93%

specificity for m/z values >1000 [14], 89% sensitivity and 92% specificity for m/z values >2000 [14]), or a unique splitting of data into one training and one testing set (98.4% overall accuracy for m/z values >2000). The external cross-validation has been proven to give unbiased estimates, and these estimates might be considered indicative of the generalization of the prediction rule to unseen data [18]. Thus, it could be argued that the proposed method should work efficiently on new ovarian cases, given that the experimental protocol used for obtaining MS-spectra remains unaltered.

Although the external cross-validation process may be used for reliable estimates regarding the performance of the system, it cannot be used to propose a specific pattern recognition design with particular m/z value. This is so since, at each one of the 10 repetitions of the external cross-validation procedure, it is most likely that a different combination of optimal design features will result for a particular classifier. Thus, it will be difficult, at the end of the 10-trial validation procedure, to propose a particular set of optimal features, which will provide a classifier design with highest classification accuracy to "unseen" MS-spectra.

One proposition for designing a system ready-to-use could be the selection of those m/z values that were simultaneously chosen by all classifiers at each repetition of the external cross-validation process. Such a selection is of value since all classifiers utilized in this study have exploited MS-spectra from a different point of view due to their different theoretical concepts: the PNN is based on non-parametric estimation of the probability density function of each class, the SVM relies on geometrical criteria, by seeking an image of the input space into a higher dimensional feature space in which data can be linearly separated, and the k-NN focus on the relationship of each data point with its neighbourhood. Thus, as most important m/z values were considered those appearing as 'best features' to all classifiers during the pattern recognition system's designing stage. However, these m/z values must be first identified in the protein level, and confirmed by traditional methods, such as RIA, ELISA, Western blot, in order to be accepted as useful biomarkers in the diagnosis of ovarian cancer.

The methodology described in the present study has been tested on a publicly available proteomic dataset, achieving the highest classification accuracy compared to other studies that have employed the same dataset. However, for our method to have a meaningful clinical contribution it should be further

tested on ovarian datasets from well-designed experiments. With regards to the dataset employed in the present study, it has been acknowledged [19,30] that the healthy and cancer samples were run at different times, thus allowing for a probable machine drift over time.

As a conclusion, the contribution of the present study is to propose a methodology, which is based on the concept of combining information from different MS-spectra intensity regions into an ensemble classification structure, in order to improve the accuracy of proteomic MS-spectra classification.

Acknowledgement

This work was supported by a grant from the General Secretariat for Research and Technology, Ministry of Development of Greece (013/PENED03) to A.B.

REFERENCES

- [1] I. Visintin, Z. Feng, G. Longton, D.C. Ward, A.B. Alvero, Y. Lai, J. Tenthorey, A. Leiser, R. Flores-Saaib, H. Yu, M. Azori, T. Rutherford, P.E. Schwartz, G. Mor, Diagnostic markers for early detection of ovarian cancer, *Clin. Cancer Res.* 14 (2008) 1065–1072.
- [2] I.J. Jacobs, U. Menon, Progress and challenges in screening for early detection of ovarian cancer, *Mol. Cell. Proteomics* 3 (2004) 355–366.
- [3] V.R. Zurawski Jr., H. Orjaseter, A. Andersen, E. Jellum, Elevated serum CA 125 levels prior to diagnosis of ovarian neoplasia: relevance for early detection of ovarian cancer, *Int. J. Cancer* 42 (1988) 677–680.
- [4] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, *Lancet* 359 (2002) 572–577.
- [5] <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>.
- [6] G. Alexe, S. Alexe, P.L. Hammer, B. Vizvari, Pattern-based feature selection in genomics and proteomics, *Ann. Oper. Res.* 148 (2006) 189–201.
- [7] A. Barla, B. Irlner, S. Merler, G. Jurman, S. Paoli, C. Furlanello, Proteome Profiling Without Selection Bias Presented at Computer-Based Medical Systems, Salt Lake City, UT, 2006.
- [8] N.O. Jeffries, Performance of a genetic algorithm for mass spectrometry proteomics, *BMC Bioinformatics* 5 (2004) 180.
- [9] J. Li, H. Liu, S.K. Ng, L. Wong, Discovery of significant rules for classifying cancer diagnosis data, *Bioinformatics* 19 (Suppl. 2) (2003) 93–102.
- [10] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, R.A. Clark, Data mining techniques for cancer detection using serum proteomic profiling, *Artif. Intell. Med.* 32 (2004) 71–83.
- [11] H. Liu, J. Li, L. Wong, A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, *Genome Inform.* 13 (2002) 51–60.
- [12] Y. Liu, Serum proteomic pattern analysis for early cancer detection, *Technol. Cancer Res. Treat.* 5 (2006) 61–66.
- [13] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, J.S. Kovach, Detection of cancer-specific markers amid massive mass spectral data, *Proc. Natl. Acad. Sci. U.S.A.* 100 (2003) 14666–14671.
- [14] G. Alexe, S. Alexe, L.A. Liotta, E. Petricoin, M. Reiss, P.L. Hammer, Ovarian cancer detection by logical analysis of proteomic data, *Proteomics* 4 (2004) 766–783.
- [15] T. Fushiki, H. Fujisawa, S. Eguchi, Identification of biomarkers from mass spectrometry data using a “common” peak approach, *BMC Bioinformatics* 7 (2006) 358.
- [16] J.M. Sorace, M. Zhan, A data review and re-assessment of ovarian cancer serum proteomic profiling, *BMC Bioinformatics* 4 (2003) 24.
- [17] Y. Yasui, M. Pepe, M.L. Thompson, B.L. Adam, G.L. Wright Jr., Y. Qu, J.D. Potter, M. Winget, M. Thornquist, Z. Feng, A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection, *Biostatistics* 4 (2003) 449–463.
- [18] C. Ambroise, G.J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, *Proc. Natl. Acad. Sci. U.S.A.* 99 (2002) 6562–6566.
- [19] K.A. Baggerly, J.S. Morris, K.R. Coombes, Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments, *Bioinformatics* 20 (2004) 777–785.
- [20] L. Andrade, E. Manolakos, Signal background estimation and baseline correction algorithms for accurate DNA sequencing, *Bioinformatics* 35 (2003) 229–243, VLSI, special issue.
- [21] H.W. Resson, R.S. Varghese, M. Abdel-Hamid, S.A. Eissa, D. Saha, L. Goldman, E.F. Petricoin, T.P. Conrads, T.D. Veenstra, C.A. Loffredo, R. Goldman, Analysis of mass spectral serum profiles for biomarker selection, *Bioinformatics* 21 (2005) 4039–4045.
- [22] W.S. Cleveland, Robust locally weighted regression and smoothing scatterplots, *J. Am. Stat. Assoc.* 74 (1979) 829–836.
- [23] X. Wang, W. Zhu, K. Pradhan, C. Ji, Y. Ma, O.J. Semmes, J. Glimm, J. Mitchell, Feature extraction in the analysis of proteomic mass spectra, *Proteomics* 6 (2006) 2095–2100.
- [24] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M.C. Hung, H.M. Kuerer, Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform, *Proteomics* 5 (2005) 4107–4117.
- [25] J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, R. Kobayashi, Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics* 21 (2005) 1764–1775.
- [26] N. Christanini, J.S. Taylor, *An Introduction to Support Vector Machines and other Kernelbased Learning Methods*, Cambridge University Press, 2000.
- [27] S. Theodorides, K. Koutroumbas, *Pattern Recognition*, 2nd ed., Academic Press, 2003.
- [28] L.I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons Inc., New Jersey, 2004.
- [29] K. Jong, E. Marchiori, M. Sebag, A. van der Vaart, Feature selection in proteomic pattern data with support vector machines, in: *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2004, p. 41.
- [30] K.A. Baggerly, K.R. Coombes, J.S. Morris, Bias randomization, and ovarian proteomic data: a reply to producers and consumers, *Cancer Inform.* 1 (2005) 9–14.