GMDH-based Approach for Analysis of Mass Spectra in Clinical Proteomics

Dimitri V. Nowicki[1,2], Vladislav Shaposhnik[3], Ali Bouamrani[1], Marie Arlotto[1], François Berger[1] and Tatyana I. Aksenova[1,3]

[1]*INSERM, U836, Grenoble, F-38043, France.*
[2] *Institute of Mathematical Machines and Systems of NASU, 42 Glushkov ave, 03187, Kiev, Ukraine*
[3] *Institute for Applied System Analysis of NASU, 37 Peremogy ave, 03056, Kiev, Ukraine*

nowicki@mail.ru, evl.evl@gmail.com, abouamrani@yahoo.fr,
marie.arlotto@ujf-grenoble.fr, fberger@ujf-grenoble.fr,
tatyana.aksyonova@ujf-grenoble.fr

Abstract. *Several architectures and algorithms of feed-forward networks and neural associative memories as well as GMDH-based polynomial NNs are tried for proteomic data analysis. The problem of chemotherapy responsiveness prediction by data of mass-spectroscopy is considered to explore potential applications of different neural paradigms for this domain.*

Keywords

Artificial neural networks, polynomial neural networks,
proteomics array

1 Introduction

Proteomics is recently emerged field of life science drastically growing at the beginning of XXI century. From clinical point of view, proteomics provides an opportunity to dispose of biomarkers for early and non-invasive detection of variety of disease as well as better prognostic and therapeutic response prediction. The rapid development of this field of nanomedicine poses wide variety of problems for data processing, data analysis and data mining ([1]). Nevertheless the state-of the art in mathematical and intelligent methods of data processing does not completely meet the demand. Majority of publications in proteomic data analysis are focused on several method of unsupervised learning and clustering (see for example [2] etc.). In the present paper we explore potential of supervised learning and recognition using an example of clinical proteomic problem. This is a task of prediction of responsiveness of the cancer of brain (gliomas) to the chemotherapy basing on data coming from mass spectrometry.

Two problems the classification of mass spectra according to responsiveness and the extraction of the markers (features) were considered. Data were preliminarily analyzed using standard methods: principal-component and linear discriminant analysis. Then several neural and neural-like paradigms in particular GMDH-type neural networks were applied in order to determine the best solver for classification problems. The best classifiers showed significantly better recognition quality than it was previously observed.

2 Data Description

A crucial issue for optimal treatment is to predict the response to the therapy. To study the problem the samples of serum of 66 patients suffered of different Glioma type (6 types) and grades (4 grades) were collected and adequate proteomic biomarkers were investigated. The proteomic arrays were extracted from the samples with ProteinChip Array. The arrays were analyzed with the Ciphergen ProteinChip Reader PCS4000 model. The mass spectra of proteins were generated by using an average of 530 laser shots. Detailed description of used SELDI-TOF techniques could be found in [3]. All mass spectra (proteomic profiles) were normalised according to the mean values and vectors of essential picks were extracted. Peak labeling was achieved using signal-to-noise ratio set to 5 for the first pass and 2 for the second pass with 0.3% of the mass window. Estimated peaks were collected.

Each data vector consists of 189 values of amplitudes for selected peaks. Data vectors were labeled according to glioma's responsiveness to chemotherapy (2 classes responsive, R=class 1 vs. non-responsive, NR=class 0). Two typical vectors of the data are shown at Fig.1. The data set contain 66 vectors: 36 of class 0, and 30 of class 1. For classification task data were split into training and test samples. Training sample contained 40 vectors (21 NR and 19 R), the independent test one – 26 vectors (15 NR+11R).

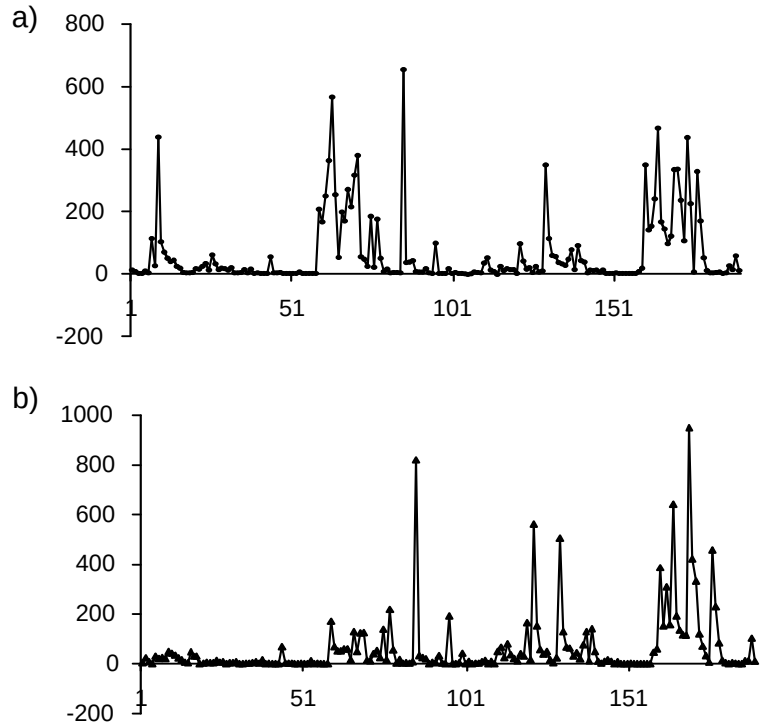

Fig. 1. Typical data vectors of classes (a) non-responsive and (b) responsive.

3 Methods

In order to know general data structure, vectors were explored using principal component analysis (Fig. 2). We can see that in low-dimensional spaces of 2-3 first principal components data have poor separability (observed degree of linear separability is 58%).

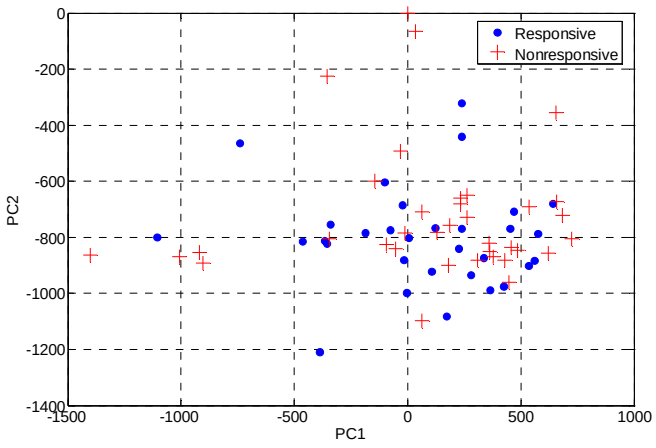

Fig. 2. Glioma data in the space of two first principal components.

3.1 Classification with ANN

We explored classification potential of the glioma data with help of different types, learning algorithms, and configurations of neural networks. For all networks input data were preprocessed using normalization to the range [0;1].

Feed-Forward Networks. We used standard multi-layered perceptron (MLP) architecture for feed-forward networks in combination with several learning algorithms. Below we present results only for networks wirh one hidden layer; augmenting hidden layer number leads only to worsening recognition quality. Three learning algorithms were selected: common backpopagation, quasi-Newtonian Davidon-Fletcher-Powell's method, and extended delta-bar-delta (EDBD) [4] algorithm. Note that small quantity of available data vectors combining with their relatively high dimension restrict choice of network architecture in order to avoid overfitting.

Associative memories. We constructed several classifiers based on neural associative memory (AM). Such networks serve usually for associative recall rather than classification; but non-iterative nature of their learning enables to avoid problem of specialization. Such thing makes associative networks especially attractive when only small set of training data is available. Among several paradigms of AM we found only two that show good results for the glioma data. They are Hamming network and kernel associative memory. Hamming network ([5],[6]) is an AM that works with real-valued data. It is able to dynamically converge during recall procedure. Roughly speaking, a Hamming AM converges to the memorized pattern nearest to an input vector.

Kernel associative memories are generalizations of Hopfield networks that use kernel-machine approach. We use an algorithm based on pseudo-inverse AM [7], there exist also some different techniques [8]. For the glioma data a hetero-associative network with Gaussian kernel was used.

3.2 GMDH-based approach, Polynomial Neural Networks

ANNs may be viewed as the universal classifiers. But the main disadvantage of this approach is that the form decision boundary is hidden within the neural network structure that makes difficulties to select the optimal feature space while the reveal of biomarkers is one of the crucial problems of clinical proteomics. Groupe Method of Data Handling (GMDH) [12] is an effective method to identify the functional structure of a model hidden in the empirical data. Iterative GMDH provided good performance in the case of high dimension of data. The main idea of Iterative GMDH is the use of feedforward networks based on short-term polynomial transfer function whose coefficients are obtained using regression technique combined with the emulation of the self-organizing activity for the neural network structural learning. Polynomial Neural Networks (PNN) ([10],[11]) belongs to the family of iterative GMDH algorithms. A twice-hierarchical neural network structure based on the

polynomial complexity control of each intermediate model [10] is used to increase the stability and computational efficiency. The vector of the number of terms and the power of the polynomial is considered as the complexity.

Robust PNN (RPNN) provides model selection and robust parameters estimation in the presence of outliers using several robust criteria for model selection and measures of goodness-of-fit. RPNN provides stable results and computational efficiency. The performance of RPNN was demonstrated with computational experiments on sixty artificial data sets where twenty percent of data were randomly distorted by large errors (outliers) [10]. The variances of outliers were *100, 200* and *300* times more than the variance of basic noise. The polynomial of forth power was successfully reconstructed and the outliers detected (Fig. 3). PNN have been successfully applied to model complex relationship hidden in the medical and pharmacological empirical data. The study of association between clinical symptoms of Parkinson's disease and patterns of neuronal activity of patients that underwent neurosurgical operation are presented in [10]. The result of the analysis allowed generating a model of dependences between clinical symptoms and neurophysiological data. An application of RPNN for the computer-aide drug design presented at [11].

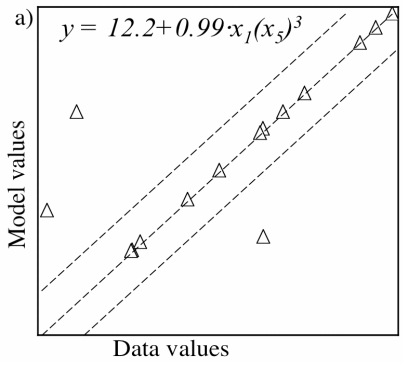

Fig. 3. Result of the computational experiment. Initial data were generated according to $y = 10 + x1(x5)3 + \xi$, where $P_\delta(\xi)= (1-\delta)\varphi(\xi)+\delta h(\xi)$; $\varphi(\xi)$ is the basic normal distribution density $N(0, 1)$; $h(\xi)$ is the distribution density of the outliers $N(0, \sigma_{out})$, $\sigma_{out}=300$; δ is the level of the large errors, $\delta=0.2$. Three outliers among fifteen observations were successfully detected with RPNN and "trough" model structure synthesized.

In present papare PNN were imploied to predict the respovsevness of the cancer of brain (gliomas) to the chemotherapy basing on data coming from mass spectrometry and to extract the biomarkers from the proteomic pfofiles.

3.3 Software

Feed-forward and AM networks were implemented in framework of multi-purpose Neuroland [12] software. PNN algorithms work with PNN Discovery Client 1.3 that presents an efficient implementation [11].

4 Results

Classification results for suitable recognition machines are shown in the Table 1. We can see that feed-forward network with EDBD learning algorithms as well as kernel associative memory can give better results on the independent test.

Tab.1. Classification results for the glioma data.

Network type	NN Configuration	Class. Rate: Training Sample, %	Class. Rate: Test Sample, %
MLP, backprop	10 hidden neurons	*100*	*65.2*
MLP, Fletcher-Powell	10 hidden neurons	*95.2*	*65.4*
MLP, EDBD	5 hidden neurons	*95.2*	*76.9*
MLP, EDBD	10 hidden neurons	*95.2*	*80.8*
MLP, EDBD	15 hidden neurons	*95.2*	*76.9*
Associative Kernel	Gaussian kernel, α=0.005	*97.5*	*80.8*
Associative Hamming	--	*100*	*61.6*
PNN	--	*100*	*61.6*
RPNN	--	*97.5*	*66.1*

Polynomial neural networks show relatively moderate classification rate but allow selecting significant features. During glioma data analysis 15 essential features were found.

5 Conclusion

The present research showed that supervised artificial neural networks are feasible and competitive for the tasks of clinical proteomics. Obtained classification rates are 15-25% better than previously known for such class of tasks (see [2] for algorithms; usually CipherGen clustering software was utilized).

This paper presents preliminary results. The volume of data is not sufficient. Moreover ANN classification gives a proof of separability of classes with degree of separability equal to 80% but does not provide the markers' extraction. Task of feature extraction and significant component selection has crucial role in the fields of clinical proteomics. We are planning to develop efficient techniques of feature selection basing on RPNNs as well as neural network's structure exploration. This will be subject of the future research.

References

[1] Belhajjame, K. et al. Proteome Data Integration: Characteristics and Challenges. *Proceedings of the UK e-Science All Hands Meeting*, ISBN 1-904425-53-4, September 2005, Nottingham, UK

[2] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* U S A 1998; 95: 14863–8.

[3] Bouamrani A, Ternier J, Ratel D, Benabid AL, Sartell JP, Brambilla E, Berger F. Direct-tissue SELDI-TOF mass spectrometry analysis: a new application for clinical proteomics. *Clin Chem.* 2006 Nov; 52(11):2103-6. Epub 2006 Sep 21.

[4] Minai, A.A.; Williams, R.D. Back-propagation heuristics: a study of the extended delta-bar-delta algorithm**;** *Neural Networks*, 1990

[5] Lipmann R.P. An Introduction to Computing with Neural Nets// IEEE AASSP Magazine. -1987. –Vol.3, №4. -P.4-22.

[6] N. Ikeda, P. Watta, M. Artiklar, M. Hassoun. A two-level Hamming network for high-performance associative memory. Neural Networks, Vol. 14 (2001), No. 1, pp. 1189-1200.

[7] D.W.Nowicki, O.K. Dekhtyarenko, "*Kernel-Based Associative Memory"* Proc. IJCNN'04, Budapest, Hungary, July 25-28

[8] Caputo, B. Niemann, H. Storage Capacity of Kernel Associative Memories *Lecture Notes in Computer Science ;* 2002, ISSU 2415, pages 51-56

[9] Madala, H.R., Ivakhnenko, A.G.: Inductive Learning Algorithms for Complex Systems Modeling. CRC Press Inc., Boca Raton, 1994

[10] Aksenova, T. Volkovich, V. Villa, A. E. P. Robust Structural Modeling and Outlier Detection with GMDH-Type Polynomial Neural Networks. *Lecture Notes in Computer Science*, 2005, NUMB 3697, pages 881-886

[11] Aksyonova, T. I. Volkovich, V. V. Tetko, I. V. Robust Polynomial Neural Networks in Quantative-structure Activity Relationship Studies. *Systems Analysis Modeling Simulation*; 2003, VOL 43; PART 10, pages 1331-1340

[12] A.M. Reznik, E.A. Kalina, A.S. Sitchov, E.G. Sadovaya, O.K. Dekhtyarenko, A.A. Galinskaya, "The multifunctional neuralcomputer NeuroLand," *Proceedings of the Int. Conf. on Inductive Simulation*, Lviv, Ukraine, vol.1 (4), pp. 82-88. May 20-25, 2002.