Bulletin of PFUR Series *Mathematics. Information Sciences. Physics.* No 3 (2). 2010. Pp. 25–31

Математическое моделирование

UDC 519.68; 633/635:577/2

Extraction of Data Features for Neuro-Classifier Input

G. A. Ososkov, D. A. Baranov

Laboratory of Information Technologies
Joint Institute for Nuclear Research
Joliot-Curie 6, 141980 Dubna, Moscow region, Russia

The problem of essential data compression to be input to ANN-classifier without loosing significant information is considered on the example of the quite substantial task of the genetic protein structure analysis, which is important for genetic biology researches in radiobiology and, especially, in agricultural. Such analysis is usually carried out by studying ElectroPhoretic Spectra (EPS) of gliadin (alcohol soluble protein) of the inspected grain cultivar. EPS digitization produces a densitogram with 4 thousands counts, which most informative features must be extracted to be input to ANN. Besides these data require special preprocessing for densitogram smoothing, pedestal eliminating, as well as compensating such digitization orocess defects as signal noise, variability of spectrum borders and illumination, their non-linear starches due to electrophoresis nonstationarity.

Several alternative approaches to features extracting were studied: (1) the densitogram coarsing into 200 averaged measurements; (2) the principal component analysis; (3) recognition of all well-pronounced peaks in order to evaluate their parameters to be input to ANN; (4)–(5) data compression by both discrete Fourier (DFT) and wavelet (DWT) transformations. These methods have been used for feature extraction from samples formed by experts for 30 different sorts. Then extracted features were used to train ANN of three-layer perceptron type. The comparative study of the recognition efficiency with data compressed by the methods listed above shows their high sensitivity to the number of sorts to be classified. Only DFT and DWT approaches could keep the efficiency on the level 95-97% up to 20 sorts.

A further development of feature extraction methods and a study of possibility to develop a hierarchy of classifying ANNs are intended.

Key words and phrases: artificial neural networks, classification, genetic analysis, electroforetic spectrum, data compression, fast Fourier transform, principal component analysis, discrete wavelet transform.

1. Introduction

Artificial Neural Networks (ANN) are widely and successfully applied to problems of classification, forecasting, and recognition. The simplicity of the structure of ANNs of MultiLayer-Perceptron (MLP) and other types stimulated many researchers to develop universal software packages that generate MLP on the basis of a specified number of layers and neurons. It is especially certain in High Energy Physics (HEP) where due to well-developed theoretical basis one has no problems with simulating the necessary sequence of data required for training a network. It is noteworthy that one of the first such neural packages, JETNET, was developed in the early 1990s by physicists at Lund University [1]. Besides at that time several firms developing electronics in cooperation with physicists succeeded in the hardware realization of widely-applicable ANNs in the form of integrated chips, which operate in parallel and allow the training of a network to an a priori simulated configuration [2].

However, for many other experimental sciences, as biology, in particular, genetics, where a reliable model of regularities may be not clear or even absent, the dimensionality of experimental data to be classified can be very high and, besides, the amount of

Received 28th November, 2009.

Authors thank Dr.A. Kudryavtsev (VIGGS) for the problem formulation and providing all experimental data and also S. Lebedev, S. Dmitrievsky, and E. Demidenko (JINR) for the essential help in performing calculations.

these data is, as a rule, scarce for training and verification of the quality of a trained network operation. Therefore, the methods of considerable ANN input data reduction without information loss and the choice of optimal network structure are of crucial importance.

In the given paper a comparative study of such methods for data reduction to be input to ANN, as the method of principal components (MPC) in its NN realization, and methods based on Fourier and wavelet filtering is given.

2. Genetic Protein Structure Analysis

It usually is fulfilled by the gel electrophoresis. It is a technique widely used in modern biochemistry and molecular biology for the separation of protein (as well as DNA or RNA) molecules by applying an electric current to a polyacrylamide gel substrate what caused to move the molecules through the gel at different rates forming a specific electrophoretic spectrum corresponding to their genetic structure. In case of breeding and seed farming in agriculture gliadin of a cultivar under study is used in the gel electrophoresis [3]. For instance, the electrophoretic spectrum of a wheat gliadin comprises 20 to 40 components differing in staining intensity. After the staining, the gels were washed with water and photographed in transmitted light, as it is depicted in fig. 1 where 17 cultivars are exposed. Each gel consists of 17 strips corresponding to different cultivars of durum wheat. One can note variability of spectra especially for outermost and middle strips due to unsteadiness of the electrophoresis process lasting for several hours.

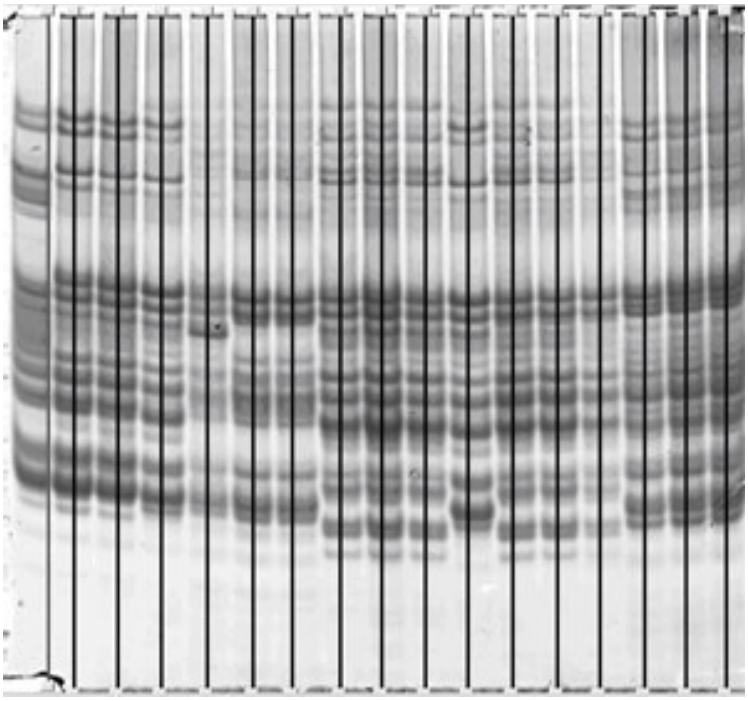

Figure 1. **Electrophoregram example**

Figure 2. **12^{th} strip of the gel image from fig. 1**

On the basis of the electrophoregram information the experts may determinate grades of wheat grains. It is tiresome work for experts, therefore it is necessary to automate this process with neural-classifiers.

3. Brief Reminder of ANN Basic Concepts

Artificial neurons are simple logical devices specified by (I) activation level; (II) topology of connections between neurons; (III) the measure of interaction with other neurons, which is referred to as synaptic coupling power or weight; (IV) output level, which is related to the activation level by a certain, usually sigmoid, function. The weights of these connections are different and can be defined in dependence of the problem under consideration. The entire system consists of a vast number of identical neurons and the result of the operation of an ANN is almost not sensitive to the characteristics of a specific neuron.

The general input signal arriving at the j-th neuron is $h_j = \Sigma w_{jk} x_k$, where x_k is the signal from the k-th neuron of the network and w_{jk} is the synaptic-connection weight. The output signal of the j-th neuron is the result of applying the activation function to this total signal, i.e. $y_j = g(o)h_j p$, where $g(o)$ is either a threshold function.

The key characteristics of a network are the type of connections between neurons and network evolution dynamics determined by the activation function for neurons and the rule of varying weights upon this evolution. Feed-forward NNs of multilayer-perceptron (MLP) type are considered further methods of training MLPs for classification and recognition. An MLP establishes a correspondence between an input vector $\bar{X} = (x_1, x_2, \ldots, x_n) = \{x_n p\}$ and an output vector $\bar{Y} = \{y_j\}$. In particular, for a three-layer perceptron, input signals are, first, transformed in the hidden-layer neurons, as $h_k = g(\sum\limits_i w_{ik} x_i)$. Then, the signals from the hidden-layer neurons are transformed by the output-layer neurons, as $y_j = g(\sum\limits_k w_{kj} h_k)$. This transformation $\bar{X} \Rightarrow \bar{Y}$ is completely described by synaptic weights $\{w_{ik}\}$; $\{w_{kj}\}$, which should be found to use an MLP for solving a particular problem.

Weights can be determined, if there is a set of data with known properties, the so-called **training sample** consisting of pairs of vectors $(\{x_i\}^{(m)}, \{z_j\}^{(m)})$, $m = 1, M$, where M is the sample size. The training of MLP by the most frequently used method of **error back propagation** is based on comparing the vectors of these pairs, i.e., the known classification result $\bar{Z}^{(m)}$ and MLP yield $\vec{Y}^{(m)}$ through the square functional:

$$E = \sum_{mj} \left(y_j^{(m)} - z_j^{(m)} \right)^2 \Rightarrow \min, \tag{1}$$

with the use of the weights $\{w_{ik}\}$; $\{w_{kj}\}$ as minimizing parameters [4]. The solution is usually sought by the method of steepest descent. Equating the derivatives of functional (1) with respect to weight parameters to zero, we obtain the iterative rules of changing weights at each training epoch.

4. Neuro-Classifier Application for the Genetic Problem

Following procedures give us the set of standardized densitometry data for each of wheat cultivars (see fig. 3), which were previously classified by experts.

Before input data to the neural network, it should be prepared. All processing consists of two stages: preprocessing and processing. Preprocessing includes following stages:

1) digitization and standardization of densitometry data;
2) denoising and eliminating background pedestal;

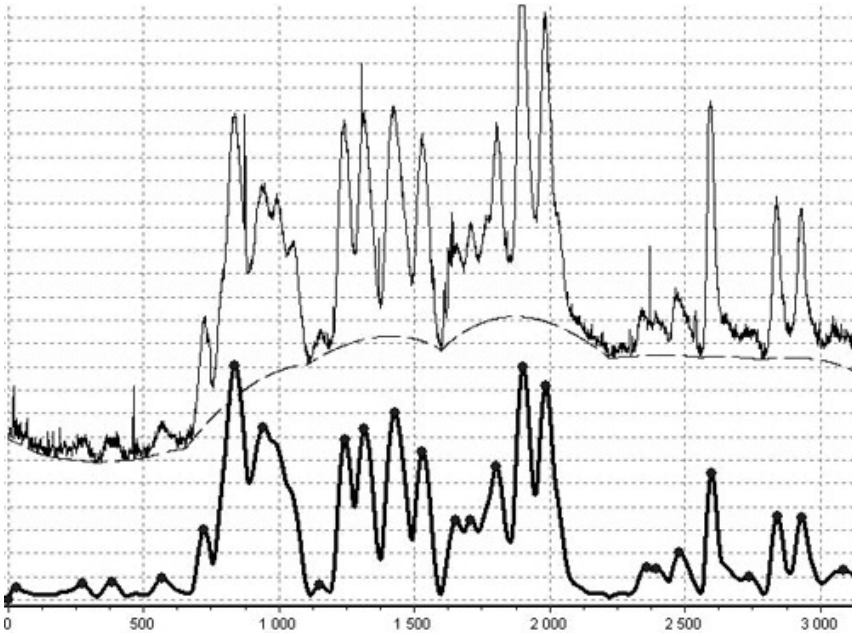

Figure 3. **Densitometry result of some of gel strips. Thin solid line — original signal with background noise and pedestal, broken line — background pedestal, thick solid line — smoothed signal without pedestal, dots — detected peaks**

3) data smoothing;
4) density normalization to the range 0–255;
5) aligning strips to fix the beginning and the end of information on the gel.

Processing includes following stages:

1) extracting the most informative features;
2) giving processed data to input of ANN.

In this work following feature extraction approaches were applied:

1) Coarsing data to 200 zones with mean density;
2) Principal Component Analysis (PCA) [5];
3) Fast Fourier Transform (FFT);
4) Discrete Wavelet Transform (DWT);
5) Ranking data by peak integral [6].

1^{st} approach: spectrum coarsing from 4000 points into 200 zones with averaged density.The real size of the training sample is 120 etalons preliminarily classified by experts for each of 20 wheat sorts, i.e. for 5 different sorts we have 600 etalons for training. Result for 5 sorts: after training ANN the efficiency was 85%.

2^{nd} approach: The size of a training sample can be efficiently reduced without essential information loss by the Principal Component Analysis (PCA) [5]. The application of an ANN for analyzing large objects such as digitized images implies the use of many thousands of neurons. In this case, the informative part of the object under analysis occupies much smaller subspace in the space of input features. The method of principal components provides the possibility of projecting vectors onto this subspace with the conservation of the substantial features of the object under analysis by using the information of their cross correlation, i.e., the covariance matrix $S_X = cov(X_i X_k)$ [7]. To this end, one applies the orthogonal Karhunen–Loeve transform $L = \{l_{ki}\}$ [8], which transforms S_X to the diagonal form, where eigenvalues λ_i are numbered in decreasing order. Thus, we can retain only m most significant eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m (m \ll p)$ and express the input data in terms of these principal components as

$$X_i \cong l_{1i} Y_1 + l_{2i} Y_2 + \ldots + l_{mi} Y_m.$$

PCA has its neural network implementations what allows to avoid cumbersome calculations of covariance matrices and their eigenvectors. It is done by so-called recircular (autoassociative) NN. Such NN uses 4000 input neurons, as etalon, output ones. The number of hidden neurons Nhid should correspond to the number of principal components. The best efficiency of the next classifying network was obtained for Nhid = 150, i.e. data was compressed in more than 20 times. Result for 5 sorts: after extracting 150 PC from all data of training and testing samples the PCA classifying efficiency was 99.54%. This efficiency keeps stable while the increase of sorts number till 8 then drops down for 17–25%.

3rd **approach**: The Fast Fourier Transform is an extremely important and widely-used method of extracting useful information from sampled signals. It is an efficient algorithm to compute the Discrete Fourier Transform (DFT) and its inverse. There are many distinct FFT algorithms involving a wide range of mathematics, from simple complex-number arithmetic to group theory and number theory; this article gives an overview of the available techniques and some of their general properties, while the specific algorithms are described in subsidiary articles linked below. An FFT computes the DFT and produces exactly the same result as evaluating the DFT definition directly; the only difference is that an FFT is much faster. (In the presence of round-off error, many FFT algorithms are also much more accurate than evaluating the DFT definition directly, as discussed below). The FFT is defined by the formula:

$$H_n = \sum_{k=0}^{N-1} h_k \exp\left(\frac{kn}{N} 2\pi i\right), \quad n \in \left[-\frac{2}{N}, \frac{2}{N}\right].$$

Real part of direct FFT was used to transform input data to the frequency domain, where the highest frequencies were cut up to 256 (16 times of reduction). After transforming all training samples to Fourier space NN-classifier (256/40/5) was trained on them and tested again on transformed sample. Result: 100% of efficiency and less for more than 5 sorts. For example, for 8 sorts it was 80–90%.

4th **approach**: The Discrete Wavelet Transform (DWT) has a huge number of applications in science, engineering, mathematics and computer science. Most notably, it is used for signal coding, to represent a discrete signal in a more redundant form, often as a preconditioning for data compression. DWT converts an input series x_0, $x_1, \ldots, x_m$, into one high-pass wavelet coefficient series and one low-pass wavelet coefficient series (of length $n/2$ each) given by:

$$H_i = \sum_{m=0}^{k-1} x_{2i-m} \cdot s_m(z), \quad L_i = \sum_{m=0}^{k-1} x_{2i-m} \cdot t_m(z),$$

where: $s_m(z)$ and $t_m(z)$ are called wavelet filters, K is the length of the filter, and $i = 0, \ldots, [n/2] - 1$.

Coiflet DWT of the 6th order were applied to transform all training and testing samples into wavelet space. Then NN-classifier (256/40/5) was trained and tested on them. Real efficiency was almost such as FTT.

5th **approach**: This method is based on the peak extraction and the peak order in which higher and lower peaks are alternating. It was proposed to recognize all well-pronounced peaks, fit each of them by some of bell-shaped function, like a Gaussian, in order to evaluate 3 basic parameters of each peak: position, integral (square under this peak) and its rank according to its integral.

The maximum number of peaks on every of all densitogramms given to us was equal to 37, so there were 111 (37×3) input neurons, 5 output and 40 hidden neurons.

Result for 5 sorts: after training ANN the efficiency is 100%. For more number of sorts, for example 8 sorts, efficiency was from 88% till 98%. Such efficiency is more then other methods have.

5. Reasons of the Classifying Efficiency Decrease

Almost all spectra have distortions and noise. It leads to a variability of spectra even in the case of the same sort. On the contrary, spectra of different, but genetically close sorts can be sometimes almost identitical (see fig. 4).

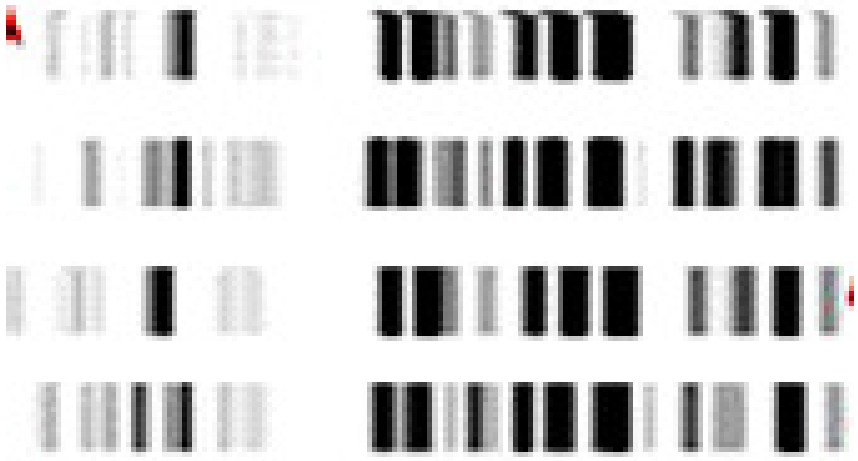

Figure 4. **Densitometry result of some strips (6th, 13th, 19th and 26th sorts)**

6. Summary

Summarizing our comparative study of five above listed methods we can conclude that although some of them show satisfactory results on classifying 5-8 sorts of wheats, further increasing sort numbers causes drop of classifying efficiency. The ranking method advantage when the sort number is growing up should be noted.

Software system was elaborated in collaboration with the VIGG RAS Institute for the full chain of electrophoretic data genetic analysis.

The further system development is thought to apply Kohonen ANN and to formalize expert classifying approaches in order to elaborate a hierarchy of ANN for the wheat protein classification.

References

1. *Peterson C. et al.*, 1993. — JETNET 3.0: A Versatile Artificial Neural Network Package. — CERN, lu tp 93-29 edition. — CERN-TH 7135/94.
2. *Lindsey C. S., Lindblad T.* Review of Hardware Neural Networks: A User's Perspective // HEP Neural Networks. — 1994. — No TRITA-FYS-9012. — Pp. 1–10. — Talk given at the Third Workshop on Neural Networks: From Biology to High Energy Physics, Marciana Marina, Elba, Italy, 26-30 1994.
3. *Ruanet V. V., Kudryavtsev A. M., Dadashev S. Y.* The Use of Artificial Neural Networks for Automatic Analysis and Genetic Identification of Gliadin Electrophoretic Spectra in Durum Wheat // Russian Journal of Genetics. — 2001. — Vol. 37, No 10. — Pp. 1207–1209.
4. *Haykin S.* Neural Networks: a Comprehensive Foundation. — N.Y., 1994.
5. *Jolliffe I. T.* Principal Component Analysis, Springer Series in Statistics, 2nd ed. — Springer, NY, 2002.
6. *Baranov D. A., Dmitrievsky S. G., Ososkov G. A.* Protein Structures Recognition using ANN // Proc. of IV Intern. Science School / TTI SFU. — Taganrog: 2008. — Pp. 126–130.

7. *Kramer M. A.* Nonlinear Principal Component Analysis using Autoassociative Neural Networks // AIChE Journal. — 1991. — Vol. 37, No 2. — Pp. 233–243.
8. *Fukunaga K., Koontz W.* Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering // IEEE Transactions On Computers. — 1970. — Vol. C-19. — Pp. 311–318.

УДК 519.68; 633/635:577/2

Выделение основных свойств данных для их ввода в нейронный классификатор

Г. А. Ососков, Д. А. Баранов

Лаборатория информационных технологий
Объединённый институт ядерных исследований
ул. Жолио-Кюри, д.6, Дубна, Московская область, 141980, Россия

Рассматривается проблема существенного сжатия данных, подлежащих вводу в классифицирующую нейронную сеть, без потери их информативности. Изложение ведется на примере задачи генетического анализа белковых структур, важной для исследований в генетической биологии, радиобиологии и особенно в сельском хозяйстве. Подобный анализ обычно проводится с помощью изучения электрофоретических спектров (ЭФС) глиадинов (спирторастворимых белков) проверяемых сортов зерновых. При оцифровке ЭФС получается денситограмма из 4000 отсчётов, наиболее информативные признаки которой и должны быть выделены для ввода в нейросеть. Кроме того, полученные данные требуют существенной предобработки для сглаживания и устранения подложки денситограммы, а также таких дефектов процесса оцифровки, как шумы, флюктуации границ и освещённости спектров и их нелинейных растяжений из-за нестационарности электрофореза.

Было изучено несколько альтернативных методов извлечения существенных признаков: (1) огрубление денситограммы до 200 усреднённых измерений; (2) метод главных компонент; (3) распознавание хорошо различимых пиков, чтобы вводить в нейросеть только их параметры; (4)–(5) сжатие данных с помощью быстрого преобразования Фурье (БПФ) и дискретного вейвлет-преобразования (ДВП). Эти методы использовались для извлечения главных признаков из множества выборок, приготовленных экспертами для 30 разных сортов, и последующего использования признаков для обучения трёхслойного персептрона. Сравнительный анализ эффективности распознавания при использовании вышеперечисленных методов показал их сильную зависимость от числа сортов, подлежащих классификации. Лишь с помощью БПФ и ДВП методов удалось удержать эффективность на уровне 95–97% вплоть до 20 сортов.

Предполагается дальнейшее развитие методов сжатия данных и возможности использовать систему многоступенчатых нейроклассификаторов.

Ключевые слова: искусственные нейронные сети, классификация, генетический анализ, электрофоретический спектр, сжатие данных, быстрое преобразование Фурье, метод главных компонент, дискретное вейвлет-преобразование.