Pergamon

Computers Math. Applic. Vol. 34, No. 7/8, pp. 677–685, 1997

0898-1221/97 $17.00 + 0.00

PII: S0898-1221(97)00168-5

Multivariate Data Analysis Based on the ω_n^k-Criteria and Multilayer Perceptron

A. YU. BONUSHKINA, V. V. IVANOV AND P. V. ZRELOV
Laboratory of Computing Techniques and Automation
Joint Institute for Nuclear Research, 141980 Dubna, Russia
ivanov@main1.jinr.dubna.su

Abstract—A comparative study of multidimensional classifiers based on the goodness-of-fit criteria ω_n^k and multilayer perceptrons (MLP) has been carried out. It is shown that MLP exhibits the "instantaneous" learning effect and improves the quality of recognition in the case of input data represented in the form of variational series. The reasons are analyzed that underlie these effects. Recommendations for joint usage of the ω_n^k criteria and of MLPs are given.

Keywords—Multivariate classifiers, Goodness-of-fit criteria, Neural network, Variational series.

1. INTRODUCTION

The primary goal of experimental data processing consists in identification of the feature events among all the events obtained in the experiment. When an event is characterized by more than one variable, the procedure applied for constructing a multidimensional classifier is not trivial. In [1], new nonparametric ω_n^k-statistics were investigated, and goodness-of-fit criteria were constructed. On their basis, a method was developed for extracting low probability multidimensional events from a background of predominant processes [2].

Artificial neural networks (ANN) for the classification of multidimensional events have been widely used in physical experiments [3]. One such problem consists of classifying individual events represented by empirical samples of finite volumes pertaining to one of the different partial distributions composing the distribution analyzed.

In the present paper a brief description of multidimensional classifiers based on ω_n^k-criteria and ANN is presented, a comparative analysis of their powers is performed, and recommendations on their joint usage are given. The results of a MLP training are analyzed for various representations of the input data, and the reasons are investigated that lead to an "instantaneous" learning effect exhibited by the neural network and to enhancement of its power, when the data are input in the form of a variational series; reduction of the number of neurons in a hidden layer without deterioration of the recognition accuracy is discussed.

2. ω_n^k-CRITERIA AND ANN

The ω_n^k-criteria are usually applied for testing the correspondence of each individual sample (event) to the distribution known *a priori*. For practical purposes it is convenient to use the

This work has been supported by the Commission of the European Community within the framework of the EU-RUSSIA Collaboration, in accordance with ESPRIT CONTRACT P9282-ACTCS.

algebraic form of the ω_n^k-statistics:

$$\omega_n^k = -\frac{n^{k/2}}{k+1}\sum_{i=1}^{n}\left\{\left[\frac{i-1}{n} - F(x_i)\right]^{k+1} - \left[\frac{i}{n} - F(x_i)\right]^{k+1}\right\}. \tag{1}$$

where $F(x)$ is the theoretical distribution function of x, $x_1 < x_2 < \cdots < x_n$ is an ordered sample, and n is the sample size [1,2].

The following procedure for extracting feature events was developed in [2] on the basis of the ω_n^k-criteria:

(a) the spectra to be analyzed are transformed, so that the contributions of dominant distributions (in most cases these are distributions of background events) from different detectors are described by a sole distribution function $F_b(x)$;

(b) each sample, composed of values pertaining to the different transformed spectra, is tested with the aid of the ω_n^k-criterion for correspondence to the $F_b(x)$ hypothesis; in this process the feature events, which do not comply with the null-hypothesis, correspond to large absolute values of ω_n^k, resulting in their clustering in the critical region;

(c) events that happen to be in the critical region are further subjected to a second test in accordance with items (a) and (b), only with the difference that now it is precisely the feature events that are collected in the admissible region; this results in additional suppression of background events in the spectra being studied.

The procedure for data handling, described above, was applied for extracting rare events in analyzing the information obtained in several experiments [2,4].

ANNs present a new paradigm of distributed parallel calculations based on the simulation of characteristic features of live neural networks. A *feed-forward* multilayer network (multilayer perceptron) is a convenient tool for constructing multidimensional classifiers [3], although its power of recognition depends critically on the choice of input data.

Such network involves an input layer corresponding to the data analyzed, an output layer dealing with the results and, also, hidden layers. A network architecture is presented in Figure 1.

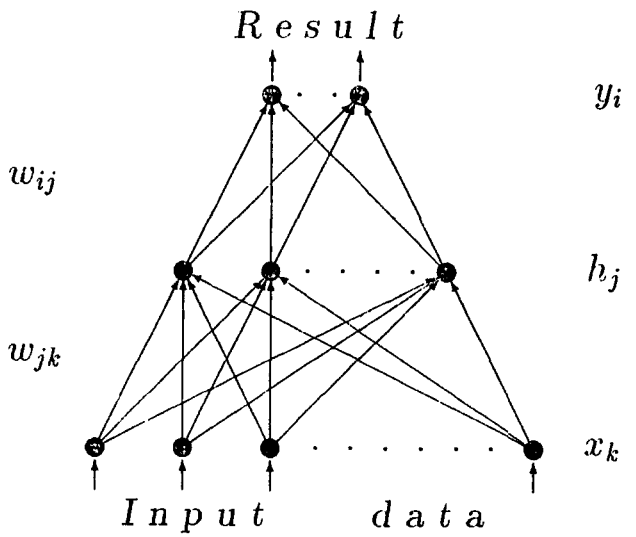

Figure 1. Architecture of multilayer perceptron with one hidden layer.

Here x_k, h_j and y_i denote the input, hidden and output neurons, respectively; w_{jk} are the weights of connections between the input neurons and the hidden layer, and w_{ij} are the weights of connections between the hidden and the output neurons. The signals $a_j = \sum_k \omega_{jk} x_k$ and $a_i = \sum_j \omega_{ij} h_j$ are fed to the inputs of hidden and output neurons, respectively. The output signals from these neurons are determined by the expressions $h_j = g[(a_j + \theta_j)/T]$ and $y_i = g[(a_i + \theta_i)/T]$,

where $g(a,T)$ is a transfer function, T is the "temperature," determining its slope, θ is the threshold of the corresponding node. Typically, $g(a,T)$ is a sigmoid, for example, of the form

$$g(a,T) = \tanh\left(\frac{a}{T}\right). \tag{2}$$

The training procedure consists in minimization of the following error functional with respect to weights:

$$E = \frac{1}{2}\sum_{p}\left[\vec{y}^{(p)} - \vec{t}^{(p)}\right]^2,$$

where $p = 1, 2, \ldots, N_{\text{train}}$ is the number of training patterns, and $\vec{t}^{(p)}$ is the desired value of the output signal.

3. DATA ANALYSIS USING ω_n^k-CRITERION AND MLP

Comparison of the powers of the indicated classifiers was carried out for the problem in which multidimensional events were generated using the Monte-Carlo method. The problem of separating cosmic protons and pions at energies of over 100 GeV was considered [5]. The ionization losses were generated for different kinds of particles traversing several detectors of an experimental setup.

The simulation of events, each of which was represented by a set of random values, namely, the energy losses experienced by the pions or by the protons in the detectors, was performed as follows. First, the kind of particle was generated, and the ratio of the pion and proton contributions was assumed to be $\pi^+ : p = 1 : 4$ (see [5]). Then, n random values ΔE_i, $i = 1, 2, \ldots, n$ (n was the number of detectors, taken to be equal to 6) were generated in accordance with the distribution of ionization losses for the selected kind of particle. The total number of generated events was set to 10000. The result of such a simulation for a single detector is presented in Figure 2; the separate contributions from protons and pions are indicated.

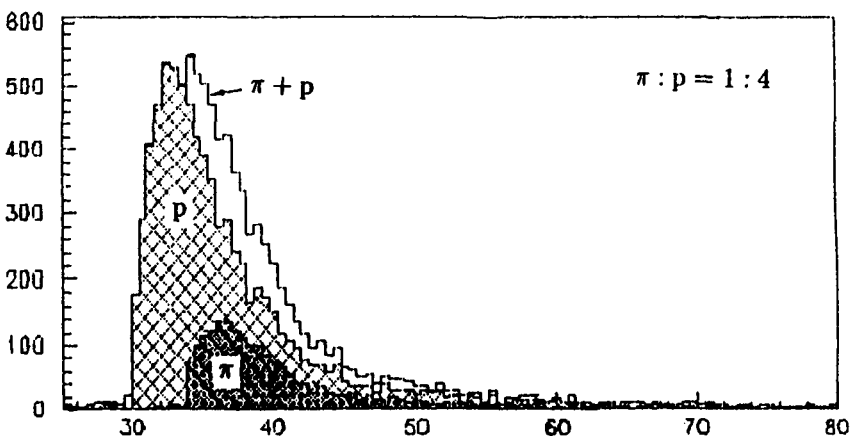

Figure 2. The resultant distribution of ionization energy losses (in KeV) for 100 GeV protons and pions in a single detector; the separate contributions from protons and pions are indicated.

The Landau distribution function was taken as the null-hypothesis in applying the ω_n^k-criterion. The following λ values were used as elements of the empirical sample:

$$\lambda_i = \frac{\Delta E_i - \Delta E_{mp}^i}{\xi_i}, \qquad i = 1, 2, \ldots, n, \tag{3}$$

where ΔE_i is the energy loss in the i^{th} counter, ΔE_{mp}^i us the transformed value of the most probable energy loss (see [2, Chapter 2]), $\xi_i \approx (1/4)$ FWHM for the distribution of ionization losses of protons in a single counter.

The application of one-sided criteria is preferable in the case of the problem considered, and since a rise in k results in an enhancement of the power of the corresponding criterion [1], the values of ω_n^k were calculated for each event by formula (1) for the maximum degree $k = 5$, for which tables of percentage points were available [1].

The distribution of the random variable ω_6^5 resulting from the processing of generated events is presented in Figure 3. The "empty" histogram is formed by proton events; the distribution for the pions is cross-hatched and is mainly located in the region of large ω_6^5.

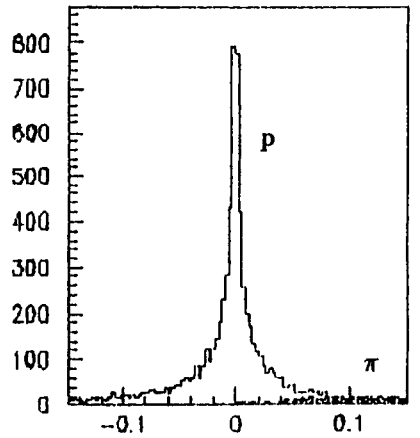

Figure 3. The distribution of the random variable ω_6^5, resulting from the processing of generated events (see text).

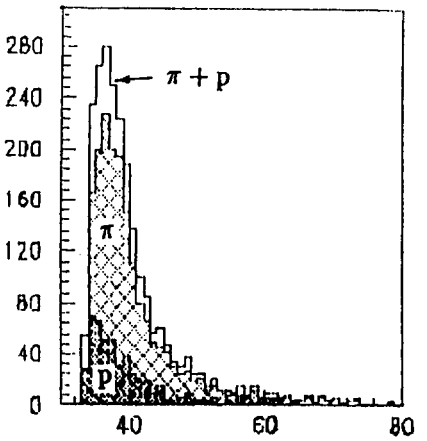

Figure 4. The distribution of energy losses for protons and pions in a single detector for selected events.

It is obviously convenient to choose the critical limit to be such that a minimum number of pion events be lost and that the contribution of proton events in the critical region be not too great. The joint distribution of energy losses for protons and pions in a single detector is presented in Figure 4 for events with $\omega_6^5 \geq 0.045$ and $\Delta E > 33.5\,\mathrm{KeV}$; the respective contributions from protons and from pions are cross-hatched. The errors of the first and second kinds, now, amounted to 6.6% and 6.4%, respectively, which practically coincided with the result of [2] obtained by utilizing the ratio of likelihood functions, which is actually the most powerful method known in the case of simple hypotheses.

A neural network containing six (in accordance with the number of detectors) input neurons, a hidden layer of 16 neurons and one output neuron was used. All the neurons had the same transfer function (2). The target signal at the network output was set equal to -1 for proton events and to $+1$ for pion events. An attempt to use a sample of generated energy losses ΔE_i, $i = 1, \ldots, n$ as input data gave no positive result. Therefore, an ordered sample of λ_i, $i = 1, \ldots, n$ values, calculated in accordance with expression (3), was adopted as the input data for the neural network, as well as for the ω_n^5-criterion.

All the generated events—approximately 2000 pion and 8000 proton events together with their respective target values—were mixed and divided into two equal (in number of events) groups. The first group was used for training the network and the second one for estimating its recognition efficiency. The identification of events was based on the value of the output signal: if it did not exceed a certain critical value, the event examined was treated as a proton event, and in the opposite case as a pion event.

For testing the network, a mixture of particles was used as input data. The spectra of output signals resulting from the processing of the generated events are presented in Figure 5: the "empty" histogram corresponds to proton events, the cross-hatched histogram to pion events. The probability for recognizing different particles in a mixture was also calculated and amounted to 93.5% (when the critical point was set equal to 0). The probabilities for recognizing proton and pion events were also calculated versus the value y_g, like in the case of the ω_6^5-criterion.

The cumulative probability $F(y) = \Pr\{y < y_g\}$ for pion events and the $1 - F(y)$ dependence for proton events are presented in Figure 6. The probabilities of errors of the first and second kinds at the intersection point of the two curves are equal to 6.8% and 6.9%, respectively, which is close to what is obtained using the ω_6^5-criterion.

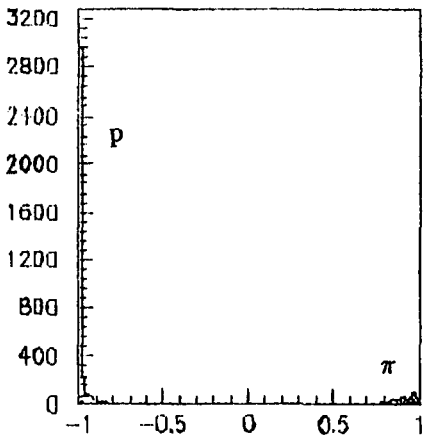

Figure 5. The spectra of output signals of the neural network: "empty" histogram—proton events; cross-hatched histogram—pion events.

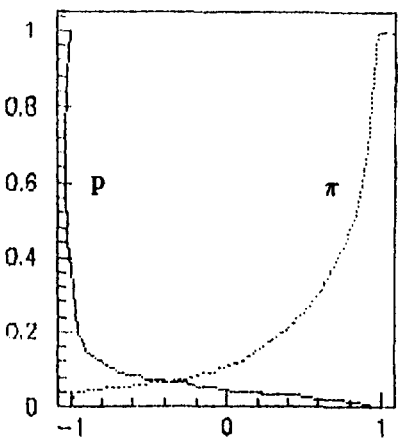

Figure 6. The cumulative probability $F(y) = \Pr\{y < y_g\}$ for pion and $1 - F(y)$ dependence for proton events: neural network.

The results may be summed up as follows:[1]

1. The presentation of input data for the MLP in the form of variational series leads to an "instantaneous" learning effect and to enhancement of its power. For passing from the source sample to the new variables, knowledge is required only of the parameters of the dominant distribution. However, at the stage of the neural network training, the information is required on the various distributions forming the experimental spectrum.
2. When only the parameters of the dominant distribution are known, goodness-of-fit criteria ω_n^k serve as a convenient tool for recognizing events corresponding to different distributions. *It must be noted that their usage is substantiated quantitatively, while the results yielded by MLP are only qualitative.*
3. The ω_n^k-criteria are convenient in that their repeated application permits extraction of the contributions of any number of partial distributions from the resultant spectrum observed in an experiment. This makes possible, for instance, upon estimation of the parameters of the constituent distributions, to additionally make use, if necessary, of a neural network.

4. THE "INSTANTANEOUS" LEARNING EFFECT

Now, let us deal with the problem of classifying events represented by samples x_i, $i = 1, \ldots, n$ of volumes from $n = 2$ to 9 pertaining to Gaussian distributions with coinciding mean values $N(0,1)$ and $N(0,0.3)$.[2] We shall consider events belonging to the distribution $N(0,1)$ to be of type I and events from $N(0,0.3)$ to be of type II.

For classifying the events we applied the *feed-forward* network, which involves n input neurons, a hidden layer of 16 neurons and a single output neuron. All the neurons have the same transition function of the form (2). In training the MLP, the output signal of the network was set equal to -1 for events of type I and $+1$ for events of type II. A total of 8000 type I events and of 2000

[1]The example examined is typical for certain problems in experimental intermediate and high energy particle physics (see, for instance, [2,4–6]).

[2]$N(a, \sigma)$ denotes the Gaussian distribution with mean value μ and variance σ^2.

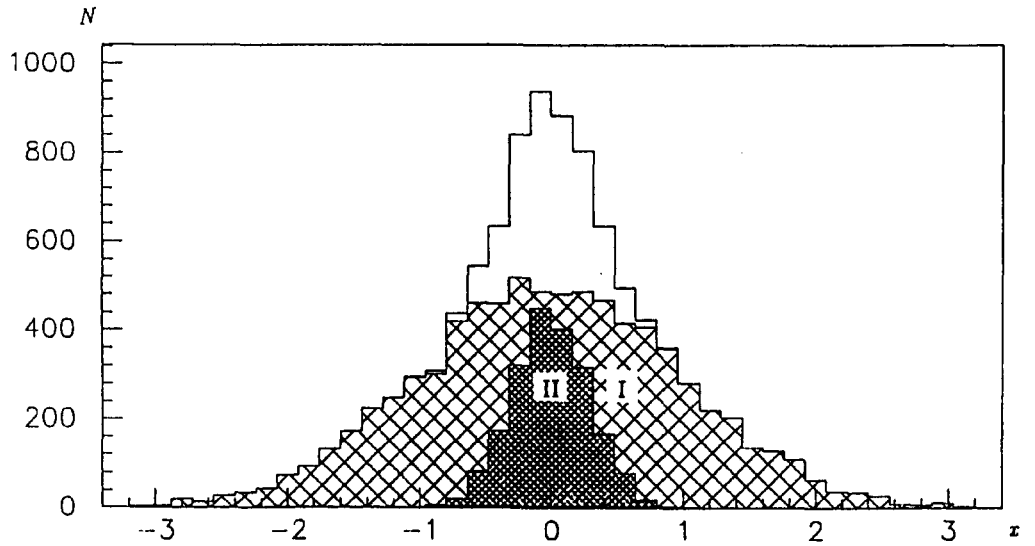

Figure 7. Total distribution for mixture of samples from the two distributions $N(0,1)$ and $N(0,0.3)$; the individual contributions from distributions I and II are, respectively, marked by appropriate shadings.

type II events were subjected to classification. The total distribution for the mixture of samples from both distributions is shown in Figure 7.

Event identification was made by the amplitude of the output signal y: if it did not exceed the given threshold value of y_t, the event examined was considered to be of type I; otherwise it was assigned to type II. In each case, the number of training cycles (epochs) required for adjustment to the actual problem being solved was realized. Here, the input data to the network were the same for each cycle, while upon its completion, correction was performed of the weights of the interneuron connections.

Now, consider the case when samples of random quantities x_i are input to the neural network. For a sample of volume $n = 2$ the network starts distinguishing events belonging to different distributions only after 485 epochs. As n increases, fewer and fewer training cycles are required for achieving a good level of recognition. Figure 8a presents an example of event recognition probability versus the number of training cycles for $n = 5$.

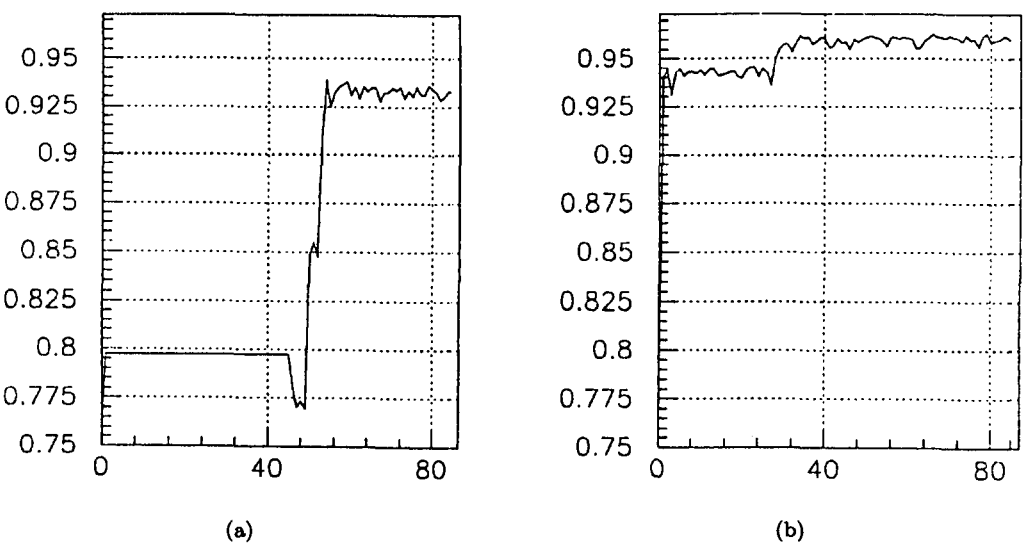

Figure 8. Probability of event recognition by the neural network versus the number of training cycles for $n = 5$, when the input data are represented: (a) by samples of random quantities, (b) by samples reduced to variational series.

We now pass from the initial sample to the ordered sample $\tilde{x}_i$ (variational series) composed of the elements x_i $(i = 1, \ldots, n) : \tilde{x}_1 < \tilde{x}_2 < \cdots < \tilde{x}_n$. Figure 8b presents the corresponding probability curve for event recognition, when ordered samples ($n = 5$) are input to the network. The nature of the curve indicates that training of the network practically takes place instantaneously. Comparison of the dependencies presented in Figures 8a and 8b reveals that the probability of event recognition being successful after completion of the training process turns out to be higher for the version involving the ordered sample.

Thus, when a variational series is utilized as the input data for the MLP, the learning time of the network is significantly reduced and its power is enhanced.

5. ANALYSIS OF REASONS OF "INSTANTANEOUS" LEARNING AND IMPROVEMENT OF RECOGNITION

In the case of Bayesian classification involving a minimum error level, separation of ν classes ω_j, $j = 1, 2, \ldots, \nu$ of events, pertaining to multidimensional Gaussian distributions with respective covariance matrices Ξ_j and vectors of mean values $\vec{\mu}_j$, is performed with the aid of separating functions of the following form (see, for example, [7]):

$$g_j(\vec{x}) = -\frac{1}{2}(\vec{x} - \vec{\mu}_j)^t \Xi_j^{-1}(\vec{x} - \vec{\mu}_j) - \frac{1}{2}\ln|\Xi_j| + \ln P(\omega_j), \quad \mu^t, \text{ transpose}, \tag{4}$$

where $P(\omega_j)$ are the *a priori* probabilities reflecting the initial knowledge of the relationship between the classes being identified, $|\Xi_j|$ is the determinant of the covariance matrix.

Two classes, ω_1 and ω_2, correspond to the problem dealt with in the present work. For simplicity we shall consider $P(\omega_1) = P(\omega_2)$, meaning the ratio between the classes to be 1:1 and permitting the term $\ln P(\omega_j)$ in expression (4) to be neglected. Moreover, the covariance matrices Ξ_j for the problem under consideration are assumed to have a diagonal structure, i.e., $\Xi_j = \sigma_j^2 I$, $j = 1, 2$, where I is the unit matrix.

The surface of solutions satisfying the condition $g_1(\vec{x}) = g_2(\vec{x})$, is determined by the equation

$$\sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i b_i + c = 0, \tag{5}$$

where

$$b_i = \frac{\sigma_2^2\mu_{1i} - \sigma_1^2\mu_{2i}}{\sigma_1^2 - \sigma_2^2}, \qquad c = \frac{\sigma_1^2}{\sigma_1^2 - \sigma_2^2}\vec{\mu}_2^2 - \frac{\sigma_2^2}{\sigma_1^2 - \sigma_2^2}\vec{\mu}_1^2 + \frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2} n \ln\frac{\sigma_2}{\sigma_1}.$$

Equation (5) defines a hypersphere with its centre at the point $(-b_1, -b_2, \ldots, -b_n)$ and of radius $R = \sqrt{-|A_1|}$, where $|A_1| = -\sum_{i=1}^{n} b_i^2 + c$. In the plane case ($n = 2$) with $\vec{\mu}_i = 0$, $i = 1, 2$, the resolving surface assumes the form of a circle with its centre at $(0, 0)$ and of radius

$$R = 2\sigma_1\sigma_2\left\{\frac{\ln(\sigma_1/\sigma_2)}{\sigma_1^2 - \sigma_2^2}\right\}^{1/2}. \tag{6}$$

Figure 9 presents the regions comprising 95% of the events composed of random samples of volume $n = 2$ pertaining to the distributions $N(0, 1)$ and $N(0, 0.3)$. The resolving circle of radius $R = 0.6901$ calculated by formula (6) for $\sigma_1 = 1$ and $\sigma_2 = 0.3$ is also indicated. Classification with respect to this boundary yields a limit recognition level of 0.8586.

Let us now consider what happens to the Bayesian boundary in the case of $n = 2$, when transition is performed from the random to the ordered sample. Ordering reduces to points $x_1 > x_2$ being transferred symmetrically about the straight line $x_2 = x_1$. In this case, the density probability of the resultant two-dimensional distribution in the region $x_2 > x_1$ becomes a doubled Gaussian density. Therefore, the expression for the resolving function does not change, and the

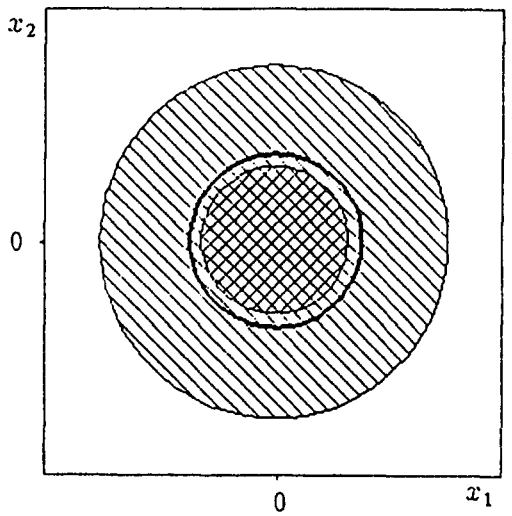

Figure 9. Confidence regions (95%) for two-dimensional distributions comprised of random samples of volume $n = 2$ from $N(0,1)$ and $N(0,0.3)$ (see text).

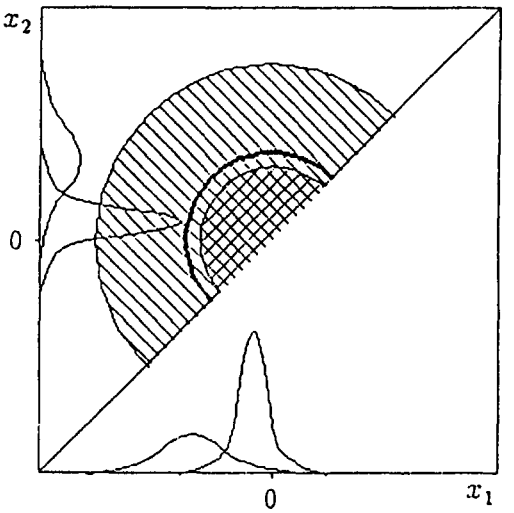

Figure 10. Confidence regions (95%) for two-dimensional distributions comprised of ordered samples of volume $n = 2$ from $N(0,1)$ and $N(0,0.3)$ (see text).

boundary is a semicircle of the same radius R (Figure 10). In both figures, the regions pertaining to different classes of event are shaded.

The transformation considered above results in the resolving boundary no longer being closed, and this significantly simplifies searching for the minimum of the corresponding functional in the course of the network training, thus enhancing the actual speed of learning. Another factor is the reduction of the "area" enclosed by the boundary, which leads to a decrease in the error during training of the network.

We now consider how the situation changes statistically in the case of transition from random to ordered samples. In this case, the elements x_m of the variational series $x_1 < x_2 < \cdots < x_m < \cdots < x_n$ correspond to different distributions with density function described by the expression [8]:

$$[B(m, n-m+1)]^{-1} [F(x)]^{m-1} [1 - F(x)]^{n-m} f(x),$$

where $F(x)$ is the distribution to which the empirical sample belongs, and $f(x)$ is its density, while the B are binomial coefficients.

Figure 10 schematically presents the density functions for the elements of ordered samples ($m = 1, 2$) pertaining to the distributions $N(0,1)$ and $N(0,0.3)$, in accordance with which events of the classes being separated are input to the neuron network. These distributions can be clearly seen to exhibit noticeable shifts with respect to the mean. These shifts can be calculated making use of the expressions for the central moments of distribution, given in [9].

For the case displayed in Figure 10, the quantities sent to the first input of the network belong to a distribution with the respective parameters $\mu_{11} = -0.5477$, $\sigma_{11} = 0.7477$ and $\mu_{12} = -0.1643$, $\sigma_{12} = 0.2243$, and the quantities sent to the second input correspond to $\mu_{21} = 0.5477$, $\sigma_{21} = 0.7477$ and $\mu_{22} = 0.1643$, $\sigma_{22} = 0.2243$. Thus, the distributions sent to each of the inputs exhibit shifts of their means by $d = 0.3834$. When transition is performed to large samples n, the value of d increases, and moreover, the divergence between the mean values of the resultant n-dimensional distributions also increases.

Thus, in the case of ordered samples, well-separated distributions are sent to the m^{th} input of the network ($m = 1, 2, \ldots, n$).[3] This feature of the transformation under study is one more

[3]In the case of disordered samples, the distributions with coinciding mean values for both classes are sent to each input of the network.

reason for the network to exhibit a significantly higher speed in "establishing" the sought resolving boundary.

It must be noted that reduction of the "area" of the separating hypersurface, besides enhancing the learning speed and the quality of recognition, also results in another important consequence consisting of the fact that the number of neurons in the hidden layer required for achieving approximately the same level of recognition is significantly lower in the case of an ordered sample. This fact becomes comprehensible if one takes advantage of the simplified model of a network involving a steplike transition function, in accordance with which each neuron corresponds to an individual hyperplane approximating part of the resolving boundary.

The resolving boundary was shown above to be a hypersphere of definite radius, when a random sample is input to the network. It can be readily shown that in the case of an ordered sample, the indicated boundary is part of the same sphere with an area $n!$ times smaller, signifying therefore that in the limit case of a sufficiently large number of approximating hypersurfaces, the number of neurons in the hidden layer, determined by the ratio between the areas of the resolving boundaries, is also $n!$ times smaller.

6. CONCLUSION

A comparative study of multidimensional classifiers based on the goodness-of-fit criteria ω_n^k and multilayer perceptrons has been carried out for distributions representing simultaneous measurements of the same physical values in several detectors of an experimental setup. It has been shown that transformation of the data, input to a MLP, into a variational series leads to significant acceleration of the network training process and, also, to improvement of the quality of recognition. Moreover, the representation of the data in such a form permits reducing the number of neurons in the hidden layer without loss of precision in the classification. Recommendations for joint usage of the ω_n^k criteria of MLPs are given.

REFERENCES

1. P.V. Zrelov and V.V. Ivanov, Neparametricheskie integral'nie statistiki $\omega_n^k = n^{k/2} \int_{-\infty}^{\infty} [S_n(x) - P(x)]^k \, dP(x)$ i ikh osnovnie svoistva. Algebraicheskaya forma, funkcii raspredeleniya i kriterii soglasiya (in Russian), JINR Communication P10-92-461, (1992).
2. P.V. Zrelov and V.V. Ivanov, The relativistic charged particles identification method based on the goodness-of-fit ω_n^3-criterion, *Nucl. Instr. and Meth. in Phys. Res.* **A310**, 623–630 (1991).
3. B. Denby, Tutorial on neural networks applications in high energy physics: The 1992 perspective, In *Proc. of II Int. Workshop on Software Engineering, Artificial Intelligence and Expert Systems in High Energy Physics*, New Comp. Tech. in Phys. Res. II, (Edited by D. Perret-Gallix), p. 287, World Scientific, (1992).
4. P.V. Zrelov, V.V. Ivanov, V.I. Komarov, A.I. Puzynin and A.S. Khrykin, Modelling of experiment on investigation of processes of subthreshold K^+—Mesons production, JINR Preprint, P10-92-369, Dubna, (1992); *Mathematical Modelling* (in Russian) **4** (11), 56–74 (1993).
5. P.V. Ramaha Murty and G.D. Demeester, *Nucl. Instr. and Meth.* **56**, 93 (1967).
6. G. Basti *et al.*, Automatic redefinition of the fuzzy membership function to deal with high fluctuating phenomena in neural nets, *INFN Roma II*, Italy, April 12, 1993.
7. R. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, (1973).
8. V.S. Korolyuk, N.I. Portenko, A.V. Skorokhod and A.F. Turbin, *Handbook of Probability Theory and Mathematical Statistics*, (in Russian), Nauka, Moscow, (1985).
9. C.E. Clark and G.T. Williams, Distributions of the members of an ordered sample, *Ann. Math. Statist.* (1958).