ELSEVIER

Neurocomputing 6 (1994) 497–512

NEUROCOMPUTING

Neural classification technique for background rejection in high energy physics experiments

Ashot Chilingarian *

Yerevan Physics, Institute, Alikhanian Brothers St. 2, Yerevan 36, Armenia

Received 16 July 1992; accepted 30 March 1993

Abstract

A comparative study of Bayesian and neural classification was done. The mathematical models of neural networks, trained in an evolutionary way, and Bayesian decision rules with Parzen-window multivariate density estimation were applied for background rejection in γ-ray astronomy experiments.

A weight function was introduced in classification score to control the relative learning 'quality' of alternative classes.

The use of a new quality function, instead of classification score, allows:

- to avoid usage of Monte Carlo events with inherent misleading simplifications and incorrectness;
- to directly optimize the desired quantity: the significance of source detection;
- to obtain the complicated nonlinear boundaries of γ-cluster.

The proposed technique can be used for background rejection in the constructing experiments of high-energy neutrino point sources identification.

Keywords: Neural networks; Multidimensional analysis; Data classification; Background rejection; Decision making

1. Introduction

The most difficult and most important part of high energy physic (HEP) data analysis is comparison of competitive hypothesis and decision making on the nature of the investigated physical phenomenon. Modern HEP apparatus consist

* Email: chilin@crd.erphy·armenia.sv

SSDI 0925-2312(93)E0017-8

of huge assemblies of electronic detectors that signal the presence and time of passage of ionizing particles or indicate the amount of charge left by a traversing particle. This bulk of diverse information on copious particles is assembled into 'events' that correspond to a single primary particle interaction.

The first decision making problem arises from the necessity of recording events on permanent storage media. The interesting new physics that is expected to embedded from the data analysis is very rare, and in contrast the already well-known noninteresting (background, noise) events are much more frequent. The signal to noise ratio can reach a value of 10^{-9}.

The so-called experimental triggers are used to make selection and reduce the amount of recorded data for subsequent analysis. In accelerator experiments, several levels (steps) of data reduction are normally used, implemented in electronics, firmware (special purpose processors) and software. Their time scale is limited by the requirement of minimization of the dead time (the time when the apparatus is unable to record events) and usually must not exceed tens of microseconds (for more details see [1,2]).

After triggering, the raw data is converted into physical variables (masses, coordinates, momentums) via the procedures of pattern recognition and estimation, and it is then recorded. Subsequent analysis may involve searching for evidence of new physics, requiring complex decision making and refined noise suppression.

So, the on-line triggering and off-line selection are the key procedures in searching for new physics and are constrained by the enormous data amounts, collection speeds and negligible signal to noise rates expected in the next generation of large accelerators (hadron colliders) known as LHC (large hadron collider, CERN, Geneva) and SSC (superconducting supercollider, Texas).

Modern air shower experiments in cosmic ray physics also are characterized by a significant increase in the volume of data collecting and therefore in the processing time to analyze this event [3]. Thus both in accelerator and cosmic ray experiments new approaches are needed which attempt to reduce the decision time and make the procedure tolerant of noise and missing data.

Another peculiarity of data analysis in high energy physics is the very intensive use of Monte Carlo simulations [4]. At any stage of the off-line analysis the simulated data are widely used; simulated data samples (training samples) are the basis of decision making on the nature of real events. [5]. Thus the proper and complete utilization of simulated data is one of the crucial aspects of data handling procedures. The Neural Network (NN) approach meets all the requirements discussed above and provides promising applications for triggering and pattern recognition at high interaction rates.

Some applications already exist in HEP: the NN method is very efficient for extracting features in hadronic data. World record performance is obtained for quark/gluon separation. The network is able to reduce the QCD background to W/Z jets by a factor of 20–30 [6,7].

Another example is connected with the most exciting discovery in experimental astrophysics of the last decade – the detection of a flux of high energy particles

from point sources. Recently, ground-based experiments have demonstrated the ability to unambiguously detect γ-rays from the Crab Nebula [8]. The Cherenkov air shower technique detects an electromagnetic cascade in the atmosphere several kilometers long and a few tens of meters wide. The characteristics of the detected shower image (length, width of the flash, the reconstructed ellipsoid axis angles with respect to the optical axis of telescope, etc.) permits rejection of the isotropic background more than two orders of magnitude.

The first successful attempts to utilize the new classification techniques encouraged the physical community to widely incorporate the neural approach in different HEP data analysis.

But further work is needed to enable NN to be properly simulated and used to improve the way the learning process is implemented.

It is hard to derive the global concept of learning from biological observations. However, we believe that in nature the brain has evolved by trial and error and that the coarse structure of the brain is determined genetically.

Our concept consists in the application of the evolutionary methods for NN training, as the most popular backpropagation method of calculation of couplings appears to be unnatural. We suppose that the structure of NN is determined by evolution and is fixed, but the synaptic strengths (couplings) are repeatedly modified in a random search way which hopefully improves the situation until a successful matrix of couplings is found.

The random search is a universal powerful methodology, akin to the trial and error method, and perhaps it forms the basis of the unpredictable efficiency of biological neural nets.

In our previous work, learning was performed in the framework of the Bayesian paradigm, by multidimensional a posteriori probability density estimation. This method was strongly dependent on the choice of a particular nonparametric method of density estimation with its free parameters, and was rather time-consuming.

The NN classifiers can be analyzed as a special class of statistical pattern classifiers which are derived from the training samples, such as Parzen-window classifiers and K Nearest Neighbor classifiers.

The NN and Parzen classifiers are trained on the same samples and, so, for the first time, we compare the two alternative classification techniques on experimental data, thus providing the continuity in development of new information technologies.

2. The nonparametric statistical inference

We shall restrict ourselves to the binary comparisons case, that is, comparisons of two, from many competing hypotheses at a time. Our example concerns a case when we want to realize the choice of one of two well-defined hypotheses – the background rejection in γ-quanta detection with the Cherenkov imaging technique.

If the statistical statement consists in the existence of an analytic distribution family (like Poisson or Gaussian), appropriate to the problem in hand, we have a prescribed parametric model. For such parametric models a well-known concept of statistical inference consists in obtaining estimates of its parameters and verifying the validity of the chosen family [9].

The classification problem is traditionally described in terms of null and alternative hypothesis, critical and acceptance regions, first and second kind errors and level of significance [10].

The best critical region is constructed by means of Likelihood Ratio (LR):

$$\mathrm{LR}(\mathbf{x}) = \frac{p(\mathbf{x}/\theta^*_\gamma)}{p(\mathbf{x}/\theta^*_{pr})}, \tag{1}$$

where $\mathbf{x}$ is a many-dimensional observable, in our case – parameters of Cherenkov flash, $p(\mathbf{x}/\theta^*_\gamma)$, $p(\mathbf{x}/\theta^*_{\mathrm{pr}})$ – are conditioned on particle type probability density functions, obtained separately for γ-images and proton images, θ^* is the Maximal Likelihood Estimate (MLE):

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{M} \ln\, p(\mathbf{x}_i/\theta), \tag{2}$$

where $\mathbf{x}_i$ are obtained from Cherenkov telescope calibration (possible only for background) or simulations (for signal events), and M is the number of calibration or simulation trials.

For almost all problems of inference, the crucial question is whether the fitted probability family is in fact consistent with the data. Usually, parametric models are chosen for their statistical tractability, rather than for their appropriateness to the real process being studied.

Of course, any statistical inference is conditioned on the model used, and, if the model is oversimplistic so that essential details are omitted, or improperly defined, then at best only qualitative conclusions may be obtained. In HEP very sophisticated models are used, completely mimicking a stochastic mechanism whereby data is generated. Such models are defined on a more fundamental level than parametric models, and provide us with a wide range of outcomes from identical input variable sets – 'labeled', or 'training' samples (TS).

Usually, for experimental physics data handling, the Likelihood Function cannot be written explicitly, and we deal with implicit, nonparametric models, for which no parametric form of the underlying distribution is known, or can be assumed.

The Bayesian approach provides the general method of incorporating prior and experimental information and formalizes the account of all the losses connected with probable misclassification and utilizes all the differences of alternative classes [11,12]. The decision problem in a Bayesian approach is simply described in terms of the following probability measures defined on metric spaces:

(a) The space of possible states of nature $\theta = (p, \gamma)$ where p, γ are indexes of alternative classes (hypotheses);
(b) The space of possible statistical decisions $\tilde{\theta} = (\tilde{p}, \tilde{\gamma})$, the decision that the examined image is caused by a primary proton or a γ-quantum;

(c) Cost (losses) measure $C_{\theta\tilde{\theta}}$ defined on the direct product of the true states and decision spaces $(\theta \times \tilde{\theta})$. At correct classification the losses are equal to zero: $C_{p\tilde{p}} = C_{\gamma\tilde{\gamma}} = 0$. If we misclassify the signal event (error of the first kind), we decrease the efficiency of the γ-event registration. If we attribute hadronic images to γ-ray ones (error of the second kind), we increase the background contamination. As we expect a significant excess of background against signal, we are interested in a strong background rejection. It is therefore not reasonable to take the symmetric loss function $C_{p\tilde{\gamma}} = C_{\tilde{p}\gamma} = 0.5$, as we did in our earlier studies concerning the cosmic-ray hadrons classification by a transition radiation detector and iron nuclei fraction determination in the primary flux [13].

(d) Event (measurement, feature) space – a set of possible results of a random experiment – image parameter samples obtained by a Monte Carlo simulation. We shall denote these samples by ω_P and ω_γ. The experimental image handling procedure parameters are determined by these samples.

(e) The prior measure $P_\theta = (P_\gamma, P_p)$. For this measure we used the uniform distribution $P_\gamma = P_p = 0.5$. In this case classification results will depend only on the available experimental information and the losses.

(f) Conditional density (likelihood function): $\{p(x/\omega_p), p(x/\omega_\gamma)\}$. The estimation of the conditional (on particle type) density on the basis of a collection of simulations (the Bayesian learning) is a typical problem in cosmic ray and high energy physics. The application of nonparametric local density estimation methods (the kernel-type Parzen estimates [14], the K-nearest-neighbors (KNN) estimates [15]) gives the best results. Our development of these nonparametric density estimates [16] makes their use in cosmic ray physics considerably more simple and increases their precision.

(g) The *a posteriori* density $p(\omega_\theta/x) \sim p_\theta p(x/\omega_\theta)$, in which the prior and experimental information is included. As we choose prior information to have a uniform distribution, the *a posteriori* density coincides with the conditional one.

Proceeding from the above definitions we can introduce the Bayesian decision rule:

$$P(\mathbf{x}/\omega_\gamma)C_{p\tilde{\gamma}} \gtrless P(\mathbf{x}/\omega_p)C_{\gamma\tilde{p}} \rightarrow \mathbf{x} \in \begin{cases} \gamma \\ p. \end{cases} \tag{3}$$

3. Learning in the feed-forward neural networks

The basic computing element in a NN is a node (neuron). A general ith node receives signals from some number of input channels (see *Fig. 1*):

$$\mathrm{IN}_i^{l+1} = \Theta_i + \sum_{j=1}^{\mathrm{NODES}(l)} J_{ij}^l \mathrm{OUT}_j^l, \quad i = 1, \mathrm{NODES}(l+1), \quad l = 1, L, \tag{4}$$

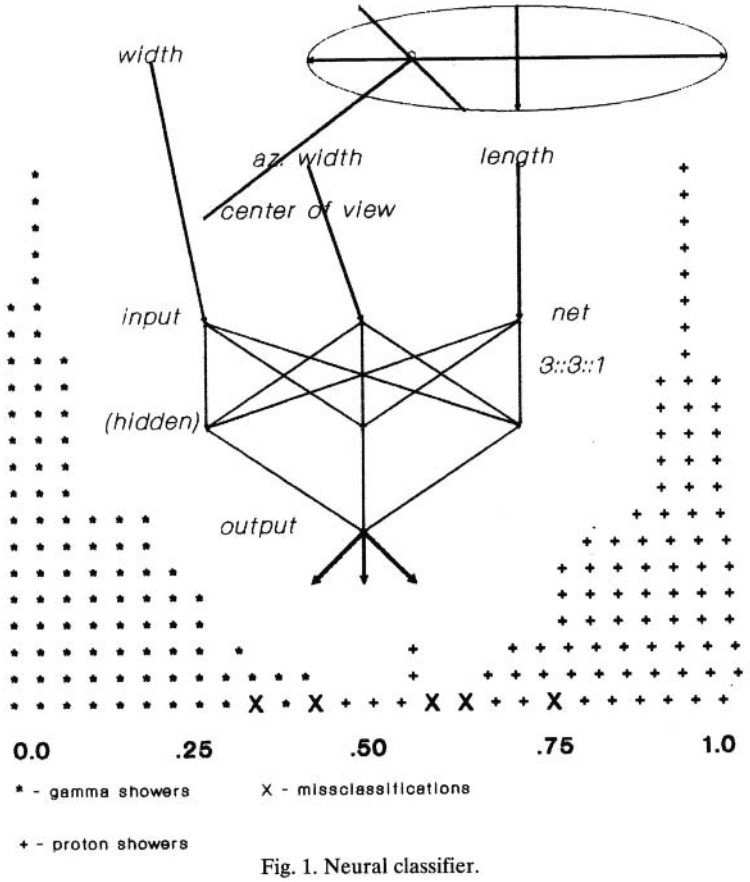


Fig. 1. Neural classifier.

where the threshold Θ_i and connection strength J_{ij}^l are parameters associated with the node i, l is the layer index, L the total number of layers, NODES(l) the neuron number in the lth layer, and OUT_j^l the output of the jth neuron in lth layer. The index j always corresponds to the higher layer (the highest layer is the input layer), and the index i to the next layer. The output of the neuron is assumed to be a simple function of this node input; usually it is formed by the nonlinear sigmoid function:

$$\mathrm{OUT}_i^l = 1/\left(1+\exp\left(-\mathrm{IN}_i^l\right), \quad i=1, \mathrm{NODES}(l), \quad 1 \neq 1, \tag{5}$$

where IN_i^l is the input of the ith neuron in the lth layer.

The topology of the NN for event classification purposes usually has a fairly restrictive form [17]:

(1) All nodes are arranged into distinct layers.
(2) The input layer has one node for each measured characteristic.
(3) The output layer has a single node, by which output the classification function is formed.

(4) The feed-forward connectivity: a node in a given layer receives input only from nodes in the preceding layer.
(5) Complete connectivity: each node in a given layer sends its output to all nodes in the next layer.

The trivial case of such an architecture is the linear discriminant function, in which the input nodes are directly connected to the output one. By implementing this topology one can obtain the best linear discriminant, proceeding from the selected input variables.

To obtain a more complicated nonlinear discriminant surface, the transition from input to output proceeds through so-called hidden layers, in which various internal representations of the input are constructed, leading to more complicated decision boundaries.

With an input/output relationship thus defined the multidimensional feature set is translated from input through hidden layers to the output node, where classification is performed. So, the NN provides one-to-one mapping of a complicated input signal to class assignments.

Such a data handling design, combining the linear summation on the nodes input, and nonlinear transformation in the nodes, allows us to take into account all distinctive information, including differences in nonlinear correlations of alternative classes of multidimensional features.

This method is easily generalized to classification with a number of classes (for example, for classification of incident cosmic radiation into 5 distinct nuclei groups). The M-node output layer can be used to separate the input stream into 2^M classes, if binary representation of the class number is used. The analog signal of the single output node can also be used to classify the events into several categories.

The 'true' output – $\mathrm{OUT}^L_{\mathrm{true}}[k]$ for kth category events is determined to maximize the shift of the alternative classes from each other:

$$\mathrm{OUT}^L_{\mathrm{true}}[k] = (k-1)/(K-1), \quad k = 1, K, \tag{6}$$

where K is the total number of classes. In the case of two classes, e.g. the first class being γ-images and the second being p-images (background), the 'true' outputs, as one can easily see, are equal to zero and one. The actual events classification is performed by comparing the obtained output value with the 'true' one.

Thus the network contains a number of free parameters: the thresholds and connection strengths, the total number of which is equal to:

$$\mathrm{NTOT} = \sum_{l=2}^{L} \mathrm{NODES}(l) + \sum_{l=1}^{L-1} \mathrm{NODES}(l)\mathrm{NODES}(l+1). \tag{7}$$

For simple net configurations, e.g. *1*::3::1, NTOT is 10, while for a 3::3::1 NTOT is 18.

The net training consists in determining these parameters using of 'labeled' events (training samples). The figure of merit to be minimized is simply the

discrepancy of apparent and target outputs over all training samples (classification score function):

$$Q = \sum_{k=1}^{K} \sum_{m=1}^{M_k} w(k)\left(\mathrm{OUT}_m^L[k] - \mathrm{OUT}_{\mathrm{true}}[k]\right)^2, \tag{8}$$

where OUT_m^k is the actual output value of training event, belonging to the kth class, and the $\mathrm{OUT}_{\mathrm{true}}[k]$ is the target value of the kth class output, where K is the number of categories and M_k is the number of events in the Kth training set, and $w(k)$ is the weight function, controlling the relative 'quality' of each class training. If we want to suppress the background contamination significantly, the background cluster must be compact and be as near as possible to the $\mathrm{OUT}_{\mathrm{true}}$ [2], i.e. to 1. So the value of $w(2)$ must be somewhat greater than $w(1)$, if the signal cluster may be much more spread near to the 0 point as 50% registration efficiency is good enough.

For feed-forward network training, a standard technique exists, providing the approximate minimization of a classification score function: the changes, initiated by the difference between the expected and predicted output pattern, have propagated back through the network.

The small correction to the network parameters (each correction associated a with particular event) are done by the steepest descent steps:

$$J_{ij}(\mathrm{new}) = J(\mathrm{old}) + \Delta_m J_{ij}, \quad \Delta_m J_{ij} = -\epsilon \partial Q / \partial J_{ij}, \tag{9}$$

where ϵ is the step size, the distance to move along the gradient, also called the 'learning coefficient'.

The couplings between the last hidden and output units are modified (after proceeding throughout the net of an event) according to

$$\Delta_m J_{1j}^L = \epsilon\left(\mathrm{OUT}_m^L - \mathrm{OUT}_{\mathrm{true}}\right)\mathrm{OUT}_j^{L-1}\mathrm{OUT}_m^L\left(1 - \mathrm{OUT}_m^L\right),$$
$$j = 1, \mathrm{NODES}(L-1), (\mathrm{NODES}(L) = 1), \tag{10}$$

where OUT_m^L is the actual response of the single output node for the mth event, OUT_j^{L-1} is the output of the jth node of the last hidden layer and $\mathrm{OUT}_{\mathrm{true}}$ is the target output of the mth event.

The couplings, connected the the hidden layers (or hidden and input layers) are obtained by the formula:

$$\Delta_m J_{ij}^{L-1} = \epsilon J_{i1}^L(\mathrm{new})\left(\mathrm{OUT}_m^L - \mathrm{OUT}_{\mathrm{true}}\right)\mathrm{OUT}_j^{L-2}\mathrm{OUT}_i^{L-1}\left(1 - \mathrm{OUT}_i^{L-1}\right). \tag{11}$$

Thus the error terms, obtained on the output, are evaluated back through the hidden layers to the input layer (hence the name backpropagation).

It is worth remembering that a steepest descent procedure cannot escape from the local minimum region of the quality function once it enters it, and the minimization is not guaranteed to converge to an absolute minimum.

Another strategy introduced recently [18,19], provides the possibility of escaping from the local minimum region and is obviously more biologically realistic. It trains

the net in an evolutionary way, implementing the procedure of trial and error for modification of the values of net parameters.

First, the particular net parameter is randomly chosen (including node thresholds as well as couplings), then the random addition (or subtraction) Δ is selected:

$$\Delta = \eta f(Q)(\mathrm{RNDM} - 0.5), \tag{12}$$

where RNDM is randomly distributed in the (0–1) interval, $f(Q)$ is the power function controlling the rate of descent when approaching the minimum and η is a random 'step' size.

If the random step is successful, i.e. the score function decreases, then the modification survives, otherwise it is subtracted and the random search procedure continues.

The iterations stop when the value of the quality function is stabilized, and no more improvements take place, thus indicating that the theoretical limit of possible classification error reduction has been reached. The resulting set of net parameters can be used for experimental data classification. We expect that data flow passing through the trained net will be divided in two clusters concentrated in the opposite regions of the (0–1) interval. Choosing an arbitrary point in this interval (the so-called decision point C^*) the classification procedure can be defined: an event whose output is greater that the decision point is attributed to the second class, while all other events belong to the first class:

$$\mathrm{OUT}^L(\mathbf{x}) \gtrless C^* \rightarrow \mathbf{x} \in \begin{cases} p, \\ \gamma, \end{cases} \tag{13}$$

where $\mathrm{OUT}^L(\mathbf{x})$ is the output node response for a particular experimental image $\mathbf{x}$.

The overlap of clusters caused by the classification errors depends on the discriminative power of the feature subset and on the learning power. By moving the decision point along the (0-1) interval we can change the relation between first and second kind of errors (the position of the decision point is the neural analog of the cost (losses) function in the Bayesian approach).

4. Statistical and neural methods of Cherenkov images classification

For a comparative study of statistical and neural classification we used 7000 events from an observation of the Crab nebula at the Whipple observatory [20] (only OFF i.f. background data were used). The training was performed with a combined TS – simulated γ – images & experimental hadron images (507 γ & 517 p events). Three features were used: LENGTH, WIDTH and AZIMUTHAL WIDTH, the best combination obtained from multidimensional correlation analysis [21].

The Bayesian procedure parameters were: the Parsen width value -0.35 for each dimension, the cost value $C_{\tilde{p}\gamma}$ ($C_{p\tilde{\gamma}}$ is equal to $1 - C_{\tilde{p}\gamma}$) varies from 0.01 to 0.1. Leave-one-out-for-a-time procedures were used to estimate classification errors.

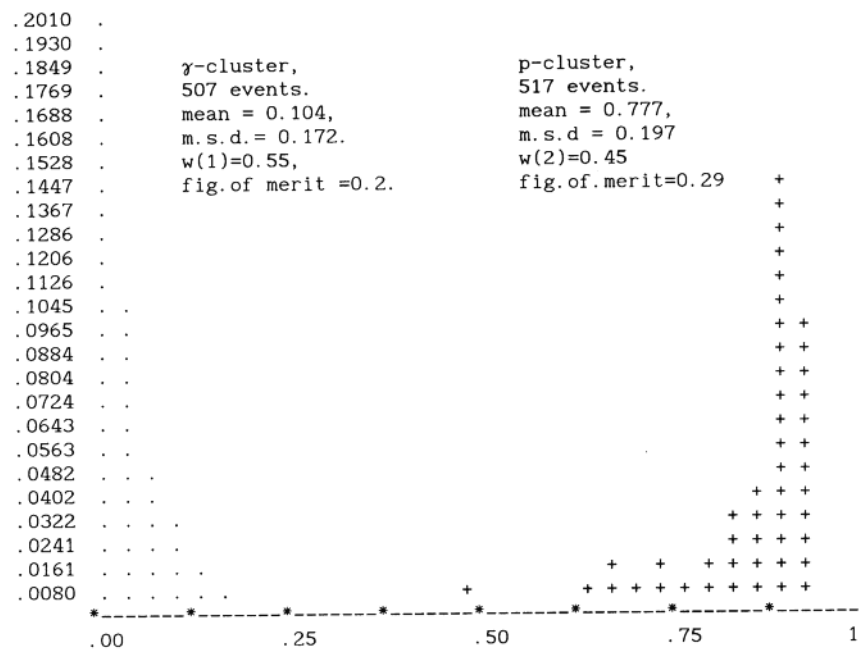


Fig. 2. The NN output after 1000 (148 success steps) trials of random search learning; tough γ-cluster; classification score improoved from 0.5 till 0.24; each sign (•, +) corresponds to 10 events.

The neural classification was performed with a simple three-layered net with 3 hidden and one output nodes. *Fig. 1* shows the Cherenkov image parameters, their input to the net and the output histograms. The decision point was varied in the same interval as the cost function in the Bayesian method. Two variants of weight values were used: 0.45 & 0.55 tight background cluster and 0.55 & 0.45 tight γ-cluster.

As one can see from the normalized histograms (*Figs. 2 and 3*) of net output after 1000 learning iterations, when couplings were modified according to (12), the alternative classes were all well-separated. The differences of the figures prove the ability of controlling the learning 'quality' of each class via altering the weights ratio (see Eq. 8). The output distribution of a well-trained class has mean values close to the target ones and relatively small spread (mean square deviation). As has been mentioned above, we are interested in significant suppression of background p-images. Therefore for data classification we use the net with couplings obtained in the learning cycle, corresponding to a 0.45/0.55 weights ratio (*Fig. 3*).

The process of successive improvement of the classification score during the random search is shown in *Figs 4 and 5*. The curves correspond to the histograms on *Figs. 2 and 3*.

From the learning results we can calculate classification errors for both the statistical and neural technique. These training sample errors (apparent errors)

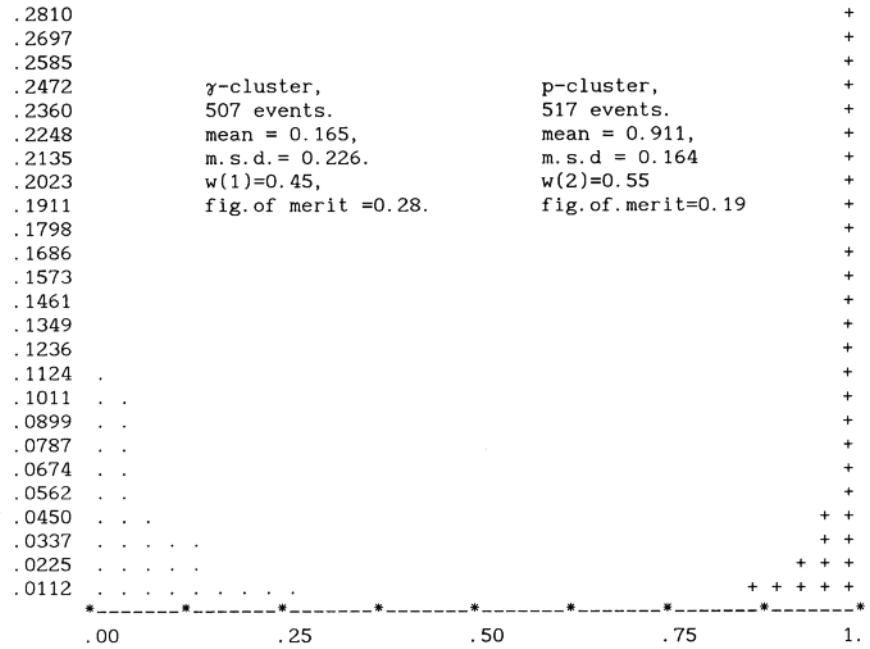


Fig. 3. The NN output after 1000 (189 success steps) trials of random search learning; tough p-cluster; classification score improved from 0.5 till 0.23; each sign (•, +) corresponds to 10 events.

usually are optimistically biased and the crucial test is, of course, the control sample test. For the control we used independent (not present in the TS) background observations (7000) events.

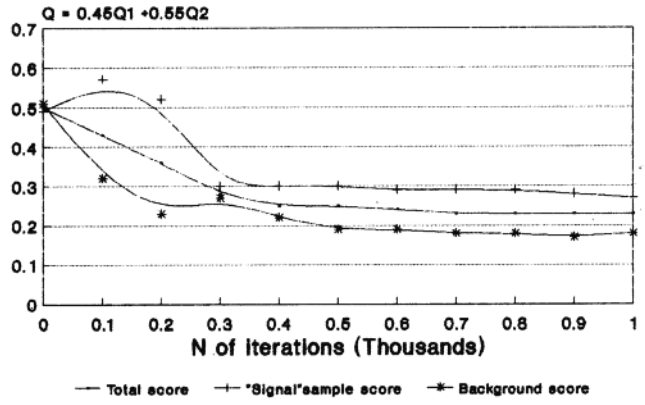


Fig. 4. Random search of net couplings, minimization of classification score.

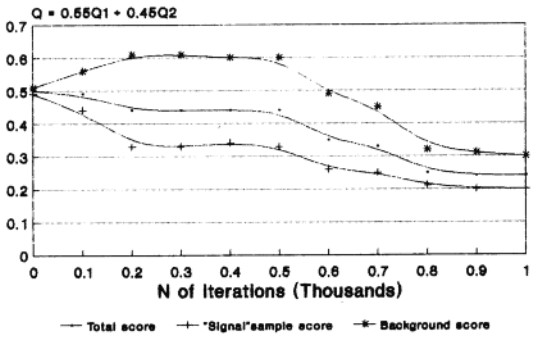


Fig. 5. Random search of net couplings, minimization of classification score.

As one can see from *Table 1*, the apparent and control errors are consistent and the statistical and neural classifications coincide very well. The lack of bias in statistical error rates is explained by use of the leave-one-out method, specially designed to avoid bias effects [22].

By changing the cost function in (3) and the decision point in (13) one can obtain different ratios between errors of the first and second kind (registration efficiency and background contamination). In this way, we obtain so-called 'influence curves' which give us the possibility of choosing the desired errors ratio in classification (remember, that you can't decrease both errors simultaneously).

The influence curves in *Fig. 6* prove again the excellent coincidence of both techniques used, that manifest the agreement of the errors to the theoretical Bayesian error limit and proper construction and learning of nonparametric statistical and neural procedures.

NN training takes ≈ 20 min. on an IBM PC/AT-386. The time spent for each experimental event classification is 0.001 sec. Bayesian training (multivariate density nonparametric estimation) is repeated for each experimental event again

Table 1
The apparent and control errors obtained by statistical and neural classifiers

	Apparent (TS) error	Control error
Bayesian classification		
γ-image registration efficiency	0.52	–
background contamination	0.006	0.0058
$C_{\tilde{p}\gamma} = 0.015$		
Neural classification		
γ-image registration efficiency	0.44	–
background contamination	0.004	0.0062
$C^* = 0.05$		

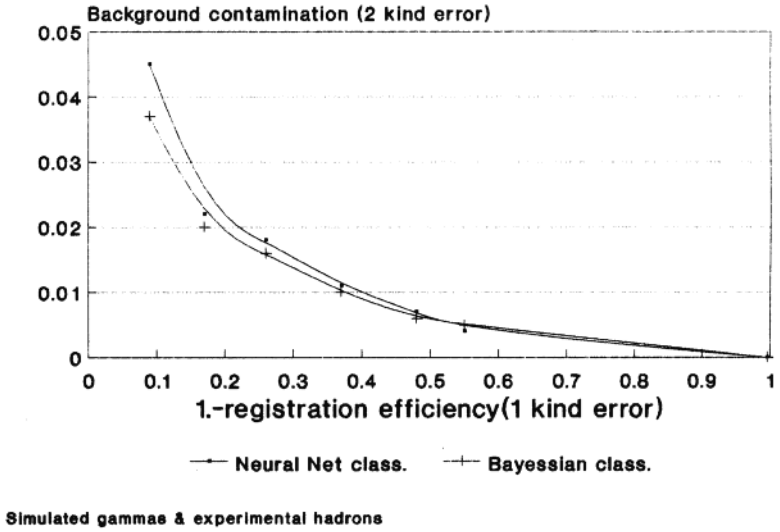


Fig. 6. Influence curve, apparent (training sample) errors.

and again, so an event classification takes ≈ 2 sec. on the same computer with same the training set and feature space.

5. The analysis of the Crab Nebula data

Searches for discrete γ-ray sources consisted in detection of an abundance $(N_{on} - N_{off})$ of events coming from the direction of a possible source comparing with a control measurement, when pure background is registered. As the expected fluxes are very weak (the signal to background ratio does not exceed 0.01), one question always has to be answered: is the detected abundance a real signal or only background fluctuation? The measure (level) of statistical significance, used in γ-ray astronomy, is the so-called criterion size (sigma) [23]:

$$\sigma = (N_{on} - N_{off})/(N_{on} + N_{off})^{0.5}. \tag{14}$$

The greater the σ value, the lesser the probability, that detected excess is due to background fluctuation, and the equipment construction and new data handling methods development have an aim to enlarge the σ value. After selecting the 'γ-like' events from row data (both from the ON and OFF samples), according to some discriminating technique, the criterion takes form:

$$\tilde{\sigma} = \left(\tilde{N}_{on} - \tilde{N}_{off}\right)/\left(\tilde{N}_{on} + \tilde{N}_{off}\right)^{0.5}, \tag{15}$$

where $\tilde{N}_{on}$ $\tilde{N}_{off}$ are the 'survived' events number.

The best discrimination technique, used in Whipple Observatory is the multidimensional cuts ('supercuts') method, proposed in [24] and then improved in [25]

(4 Cherenkov image parameters were used). This method consists in *a posterior* selection of the best γ-cluster, the multidimensional hypercube, containing 'γ-like' events. The particular coordinates of the cube were selected to maximize the σ value on the 1988–1989 Crab Nebula observation database (65 ON, OFF pairs $\sim$ 1 mln. events) [26]. Implementing the supercuts method the initial σ value was enlarged from 5 to 34.

However, it is doubtful, that the rectangular shape is the best one, and furthermore Cawley finds significant differences between ON and OFF distributions within the hypercube. We use a simple 4::5::1 neural net to select the better nonlinear shape of the γ-cluster. A preprocessing was done: events falling in the enlarged hypercube (1.5 times larger than the best ones) were selected. The net was trained on these ON/OFF events. The new quality function was used, instead of classification score (8): the sigma value (15) was maximized.

After several hours of random search the better γ-cluster was outlined and the σ value was enlarged from 27 to 36 (the record value).

The use of new quality function allows one to:

- avoid usage of Monte Carlo events with inherent misleading simplifications and incorrectness;
- direct optimize the desired quantity: the significance of source detection;
- obtain the complicated nonlinear boundaries of γ-cluster.

This modification of the neural classification method seems to be very promising for the future proton colliders (LHC and SSC) data analysis, intended to detect very rare, as yet unseen physical phenomena [27]. The effects of model dependence of the training are the main obstacles of using standard training with the classification score as quality function. The direct comparison of pure background and an mixture, containing a very small percentage of interesting events, can lead to the discovery of new physical processes and particles.

6. Conclusions

The NN classifier forms a special type of statistical classifier and is consistent with other nonparametric classifiers developed within the Bayesian approach.

After a few hundred successful random adjustments of the net parameters, the net is trained to suppress the background contamination down to a desired level, consistent with other sophisticated and time-consuming nonparametric methods.

The great advantage of NN classifiers consists in separation of the learning and the active phase. After the learning phase, all distinctive information contained in the training samples is mapped on the net parameter set. This parameter set can be written to a VLSI neurochip with the aid of a special trainer box [28]. Trained neurochips may be 4 to 5 orders of magnitude faster than software statistical classifiers. Thus we can recommend the NN classifier as a very fast intelligent trigger for on-line analysis in collider and cosmic ray physic experiments and as a sophisticated but also very fast tool for complicated off-line analysis.

The most challenging problem of modern cosmic ray physics seems to be the

identification of high energy neutrino point sources. A recently proposed approach [30] for constructing an above-ground neutrino detector (registering the upward going muons produced by charged interaction of parent neutrino in the rock below the detector), requires a very high rejection power (up to 10^{11}) to reduce the background due to upward going muons which are produced by cosmic ray interaction in the atmosphere. The proposed classification technique can be used to enlarge the sensitivity of the detector to very weak neutrino fluxes.

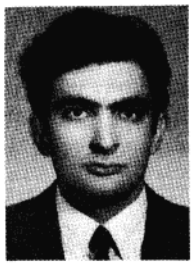
Acknowledgement

The author thanks the Whipple Collaboration for use of the Crab Nebula database and Dr. Michael Cawley for numerous useful discussions and valuable remarks concerning new methods of data analysis.

References

[1] T. Nash, High energy physics experiment triggers and the trustworthiness of software, FERMILAB-Conf-91/270, Fermi National Accelerator Laboratory, 1991.
[2] L. Gupta, B. Denby et al., Neural network trigger algorithms for heavy quark event selection in a fixed target high energy physics experiment, FERMILAB-Pub-91/117, Fermi National Accelerator Laboratory, 1991.
[3] J.C. Perret and J.T.P.M. van Stekelenborg, The applications of NN in the core location analysis of EAS, *J. Phys. G: Nucl. Part. Phys.* 17 (1991) 1291-1302.
[4] L. Lönnbland and A. Nilsson, The MC^{++} Event generation toolkit, DESY 91-158, LU TP 91-35, University of Lund, Deutsches Electronen-Synchrotron, 1991.
[5] A.A. Chilingarian, Statistical decisions under nonparametric a priori information, *Comp. Phys Comm.* 54 (1989) 381–390.
[6] P. Bhat, L. Lönnbland et. al., Using NN to identify jets in hadron-hadron collisions, Lu TP 90-13, DESY 90–144, University of Lund, Deutsches Electronen-Synchrotron, 1990.
[7] C. Petersen, Using NN to identivy the origin of gets, LU TP 90-14, University of Lund, 1990.
[8] C.W. Akerlof, M.F. Cawley et al., *Proc. 22 Internat. Cosmic Ray Conf.*, OG-sessions, Dublin (1991) 456–459.
[9] E.A. Eadie, D. Drijard, F.E. James, M. Ross and B. Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam, 1971).
[10] S. Zacks, *The Theory of Statistical Inference* (Wiley, New York, 1977).
[11] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory* (MIT Press, Cambridge, MA, 1978).
[12] J.D. Hey, *An Introduction to Bayesian Statistical Inference* (Martin Robertson, 1983).
[13] A.A. Chilingarian, and H.Z. Zazyan, On the possibility of investigation of the mass composition and energy spectra of PCR in the energy range 10^{15}–10^{17} ev using EAS data, *Il Nuovo Cimento* 14C (6) (1991) 555–566.
[14] E. Parzen, On estimation of a probability density function and mode, *Ann. Math. Stat.* 33 (1962) 1065–1087.
[15] D.O. Lofsgaarden and C.D. Quesenberry, A nonparametric estimate of a multivariate density function, *Ann. Math. Stat.* 36 (1965) 1049–1063.
[16] A.A. Chilingarian and H.Z. Zazyan, A bootstrap method of distribution mixture proportion determination, *Pattern Recognition Letters* 11 (1990) 781–785.
[17] T.D. Gottshalk and R. Nolty, Identification of physical processes using NN classifier, DOE REPORT CALT-68-160, California Institute of Technology, 1990.

[18] S. Bornholdt and D. Graudenz, General asymmetric neural networks and structure design by genetic algorithms, Desy 91-046, Deutches Electronen-Synchrotron, 1991.
[19] A.A. Chilingarian, Neural net classification of the γ- and p-images registered with atmospheric Cherenkov technique, *Proc. 22 Internat. Cosmic Ray Conf.*, OG-sessions, Dublin (1991) 540–543.
[20] M.G. Lang, C.W. Akerlof, M.F. Cawley et al., Tev observation of the Crab nebula and other plerions in the epoch 1988–91, *Proc. 22 Internat. Cosmic Ray Conf.*, Dublin (1991) 204–207.
[21] F.A. Aharonian, A.A. Chilingarian et al., A multidimensional analysis of the Cherenkov images of air showers induced by very high energy γ-rays and protons, *Nuclear Instruments & Methods* A302 (1991) 522–528.
[22] A.A. Chilingarian and S.Kh. Galfayan, Calculation of Bayes risk, *Stat. Problems of Control* 66 (Vilnius, 1984) 66–77.
[23] S.N. Zhang and D. Ramsden, Statistical data analysis for γ-ray astronomy, *Exp. Astronomy* 1 (1990) 145.
[24] A.A. Chilingarian and M.F. Cawlew, Application of multivariate analysis to atmospheric Cherenkov imaging data from the Crab nebula, *Proc. 22 ICRC*, Vol. 1 Dublin (1991) 460.
[25] M. Punch, C.W. Akerlof, M.F. Cawley, et al., Supercuts: an improved method of selecting gamma-rays, *Proc. 22 ICRC*, Vol. 1 Dublin (1991) 464.
[26] G. Vacanti, M.F. Cawlew et al., Gamma-ray observations of the Crab nebula at Tev energies, *Ap. J.* 377 (1991) 467.
[27] B. Denby, Tutorial on neural network applications in high energy physica: A 1992 perspective, Fermilab-Conf-92/121-E.
[28] C. Lindsdey, B. Denby and H. Haggerty, Real time track finding in a drift chamber with a VLSI neural network, FERMILAB-PUB 92/55, Fermi National Accelerator Laboratory, 1992.
[29] MINI Collaboration, Rejection power of a horizontal RPC telescope for left and right coming cosmic muons, Inst. Naz. Fis. Nucl. [RAPP.] AE-18, 1992.



Ashot Chilingarian is the director of the cosmic ray department of the Yerevan Physics Institute. He received the B.S. in experimental physics in 1971 at Yerevan State University, the Ph. D. degree (1984) and Doctor of Science degree (1991) at Yerevan Physics Institute. He is a member of INNS.

He has published more than eighty papers in the areas of his research interests, which include experimental cosmic ray and high energy physics, statistic methods of data analysis, multivariate density nonparametric estimation and dimensionality analysis. The main emphasis of his present research lies in the field of development of new information technologies for high energy physics data analysis, with particular emphasis on recurrent and feed-forward neural networks simulation.