ELSEVIER

April 1995

Pattern Recognition Letters 16 (1995) 333–338

Pattern Recognition Letters

Detection of weak signals against background (noise) using neural network classifiers

Ashot Chilingarian [*]

Yerevan physics institute, Alikhanian Br. 2, Yerevan 36, Armenia

Received 5 May 1993; revised 10 November 1994

Abstract

We introduce a new neural classification technique for background rejection in high energy physics and astrophysics experiments, which permits us to (i) directly optimize the desired quantity, i.e. the significance of signal detection (signal to noise ratio); (ii) obtain the complicated nonlinear boundaries for signal event acceptance. Examples of implementing the proposed technique for background rejection in high energy astrophysics experiments are presented.

1. Introduction

In high energy physics and astrophysics experiments the most important task is to separate experimental data into two classes, i.e. the signal (new interesting physical phenomenon) and background (non-interesting, abundant events). Typically in a physical experiment a particular event is described as a point in an N-dimensional measurement metric space and a mixture probability density function can be defined.

In a previous paper we considered the problem of distribution mixture classification in the case of the nonparametric type of a priori information (the statistic model is given in the form of a stochastic mechanism, whereby the data are generated and the underlying log-likelihood function cannot be given explicitly) (Chilingarian and Zazyan, 1990). A method of the distribution mixture coefficient estimation was proposed, based on the Bayesian decision rules and bootstrap replicas. The nonparametric probability density estimates used in Bayesian decision rules were obtained with the help of sets of "pseudoexperimental" events obtained in Monte Carlo simulations (training samples).

The normal approach to high energy physics data analysis is to perform precise simulations of particle collisions, determine the detector's response to the passage of the produced particles and to simulate the secondary interactions and showers (Flugge, 1991). But, if one is searching for new, yet unseen phenomena, the simulation can be misleading, and only experimental information can prove the existence of "new physics". Considering that the rate of interesting events expected in colliders of the next generation and astrophysics experiments is negligible as compared to noninteresting (background) events, we can state that the reduction of the data volumes in such a way, that the maximum sensitivity to the new physics is preserved and the maximum immunity to noise is achieved, is the key to the successful analysis (Mapelli, 1991).

[*] Email: chilin@crd.erphy.armenia.su.

0167-8655/95/$09.50 © 1995 Elsevier Science B.V. All rights reserved
SSDI 0167-8655(94)00107-3

Such "cuts" (mostly linear) performed on multidimensional distributions of measurements are usually used in collider experiments analysis to enlarge the significance of statistical conclusions concerning the existence of new phenomena. First of all it is necessary to prove with high significance level the existence of any signal, i.e. that in the mixture distribution:

$$P(x) = \alpha\, P_{\text{signal}}(x) + (1-\alpha)\, P_{\text{background}}(x) \quad (1)$$

(where usually $\alpha < 0.01$ and $P_{\text{signal}}(x)$ is unknown), α is not equal to 0. Then the value of the parameter α has to be estimated.

The fundamental theory of particle physics called the Standard Model predicts almost all of the properties of elementary particles except their masses (so we cannot obtain probability distributions of pure signal events). Determination of the masses of the top quark and the neutrino is currently the highest priority in high energy physics. After an accelerator run during 1992–1993 the D0 collaboration at Fermilab (USA) collected a data file consisting of 13 million events and one cannot expect more than a few dozen events to be the result of top quark production (D0 Collaboration, 1994).

It is difficult to outline the desired "best" signal domain where it is possible to detect the significant abundance of signal events over the background distribution — signal detection problem. The signal domain — multidimensional decision surface — can be nonlinear and its selection without knowing the signal and background distribution shapes is an unsolved problem yet.

We propose to use Neural Networks (NN) for the a posteriori signal detection. The NN technique is widely used in high-energy physics experiments for classification and event reconstruction purposes (Fogelman Soulie, 1992; Peterson and Gvaldsson, 1991; Denby, 1992). The net is usually trained on simulation data, and the so-called "classification score" is used as the objective (quality) function in minimization (best net parameters selection). It is assumed that both background and signal samples are available. But in the case of interest, the pure signal samples obtained from Monte Carlo trials are often either simplified or incorrect, so it will be better not to use "signal" samples at all. The peculiarity of our approach consists in training the net without using pure signal samples. We propose to use mixed signal and background and pure background samples obtained from the experiment for neural net training. A new type of objective function is introduced instead of the classification score.

In Section 2 the NN classification technique is presented. Section 3 describes modification to the technique and uses as an example the detection of very high energy γ-quanta coming from point sources, registered by imaging Cherenkov telescopes (Lang et al., 1991).

2. The neural classification technique

The basic computing element in a multilayered feed-forward NN is a node (formal neuron). A general ith node receives signals from all neurons of the previous layer:

$$\text{IN}_i^{l+1} = T_i + \sum_{j-1}^{\text{NODES}(l)} W_{ij}^l \times \text{OUT}_j^l,$$
$$i = 1,\ldots,\text{NODES}(l+1),\ l = 1,\ldots,L-1 \quad (2)$$

where the threshold T_i and connection strength W_{ij}^l are parameters associated with the node i, l is the layer index, L is the total number of layers, NODES(l) is the number of neurons in the lth layer and OUT_j^l is the output of the jth neuron in the 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:

$$\text{OUT}_i^l = 1/(1+\mathrm{e}^{-\text{IN}_i^l}),$$
$$i = 1,\ldots,\text{NODES}(l),\ l = 2,\ldots,L \quad (3)$$

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

With this defined input/output relationship, the multidimensional feature set is translated from input through hidden layers to the output nodes, where classification is performed. So, the NN provides the mapping of a complicated input signal to the 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 the whole distinctive information, including differences in nonlinear correlations of alternative classes of multidimensional features.

The "target" output $\mathrm{OUT}^{\mathrm{target}}(k)$ for events of the kth category (we restrict ourselves to networks with a single output node) is determined to maximize the separation of the alternative classes from each other:

$$\mathrm{OUT}^{\mathrm{target}}(k)=\frac{k-1}{K-1},\quad k=1,\dots,K \tag{4}$$

where K is the total number of classes. In the case of two classes, i.e. signal and background events, the "target" 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 "target" one. We expect, that the 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 appropriate point in this interval (the so-called decision point C^*), the classification procedure can be defined: an event with an output greater than or equal to the decision point is attributed to the background class, while all the other events are assigned to the signal class:

$$\mathrm{OUT}(x)\begin{cases} < C^*, & x \text{ is classified as signal,} \\ \geqslant C^*, & x \text{ is classified as background,}\end{cases} \tag{5}$$

where $\mathrm{OUT}(x)$ is the output node response for a particular experimental measurement x. This decision rule is a Bayesian decision rule; therefore the output signal of a properly trained feedforward neural net is an estimate of the a posteriori probability density (Ruck et al., 1990).

The expected minimal classification error caused by the overlap of the distributions (the Bayes error) depends on the discriminative power of the feature subset selected and on the learning power. By moving the decision point along the (0, 1) interval we can change the relation between the errors of the first and second kind (the position of the decision point is the neural analog of the loss function in the Bayesian approach).

The net training consists of determining the neuron couplings using the "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):

$$Q=\sum_{k=1}^{K}\sum_{m=1}^{M_k}\left(\mathrm{OUT}_m(k)-\mathrm{OUT}^{\mathrm{target}}(k)\right)^2 \tag{6}$$

where $\mathrm{OUT}_m(k)$ is the actual output value for the mth training event, belonging to the kth class, and $\mathrm{OUT}^{\mathrm{target}}(k)$ is the target value for the kth class output, where K is the number of categories and M_k is the number of events in the kth training set.

Our goal is to change the training procedure to avoid the usage of the "signal" sample. In the next section we shall introduce a new type of quality function to perform the best signal domain selection (signal detection).

3. The new neural algorithm for background rejection

In high energy astrophysics during the 1980's, significant progress has been made in the unambiguous detection of the Crab nebula at TeV energies by the Whipple collaboration (Lang et al., 1991). At TeV energies gamma rays have been shown to possess such characteristics of the Cherenkov image shape and orientation which permit them to be isolated from a much larger hadronic background. The main technique to reject this huge background consisted in applying multidimensional linear cuts on measured Cherenkov image parameters, first introduced by Hillas (1985).

To search for discrete gamma-ray sources, one looks for an abundance $(N_{\mathrm{on}}-N_{\mathrm{off}})$ of events coming from the direction of a possible source (N_{on}) as compared with the control measurement, when pure background is registered (N_{off}). As the expected fluxes are very weak (the signal to background ratio not exceeding 0.01), one should always answer the following question: is the detected abundance a real signal or only a background fluctuation? The measure (level) of statistical significance used in gamma-ray astronomy is the so-called criterion size (σ) (Zhang and Ramsden, 1990):

$$\sigma=\frac{N_{\mathrm{on}}-N_{\mathrm{off}}}{\sqrt{N_{\mathrm{on}}+N_{\mathrm{off}}}}. \tag{7}$$

Table 1
Whipple Crab detection, 1988–1990

	N_{on}^*	N_{off}^*	σ	DIFF	DIFF$/N_{\text{off}}^*$	$N_{\text{off}}^*/N_{\text{off}}$
Raw	506255	501408	4.8	4847	0.01	
Azwidth	14622	11389	20.4	3233	0.28	0.0227
Wedge cut [a]	6017	3381	27.2	2636	0.78	0.0067
Supercut [b]	4452	1766	34.3	2686	1.52	0.0035
Neural 4 : : 5 : : 1	6278	2858	35.8	3420	1.20	0.0057

[a] (Chilingarian and Cawley, 1991).

[b] (Punch et al., 1991).

The greater σ, the lesser the probability that the detected excess is due to a background fluctuation. Development of new detector hardware and new data handling methods aim to enlarge the value of σ. After selecting the "gamma-like" events from raw data (both from the ON and OFF samples), the criterion takes the form:

$$\sigma = \frac{N_{\text{on}}^* - N_{\text{off}}^*}{\sqrt{N_{\text{on}}^* + N_{\text{off}}^*}} \tag{8}$$

where N_{on}^*, N_{off}^* are the numbers of events surviving data selection cuts.

The best discrimination technique used in the Whipple Observatory is the multidimensional cuts (*supercuts*) method proposed in (Chilingarian and Cawley, 1991) and then improved in (Punch et al., 1991) (four Cherenkov image parameters were used). The method consists of a posteriori selection of the best gamma-cluster (multidimensional box), containing "gamma-like" events. The particular coordinates of the box were selected to maximize the σ value on the 1988–89 Crab nebula observation data base (65 ON, OFF pairs $\sim 10^6$ events) (Vacanti et al., 1991). By implementing the supercuts method, the initial σ value was enlarged from 5 (raw data) to 34.

For Neural Net analysis, we used the same variables as for Supercut analysis, i.e. Width, Length, Dist, and Alpha (Width and Length specify the angular size of the image, Dist specifies the position of the centroid of the image relative to the source position, and Alpha specifies the orientation of the major axis of the image relative to the source position). We use a simple 4 : : 5 : : 1 neural net to select a more realistic nonlinear shape for the gamma cluster. The net was trained on experimental ON and OFF events (i.e. events taken in the direction of the source and events taken on a background or "control" region).

A new objective function was used: instead of the classification score (6) the σ value (8) was maximized. During the iterations each particular ON & OFF event was classified according to decision rule (5) with a prechosen (or also optimized) decision point C^* and after executing all training events a new σ value was calculated. After several thousands of iterations the more complicated gamma-cluster shape was outlined and the σ value was enlarged up to 35.8 (the minimization was performed on a subsample of data, containing 25% of all events, and the σ value extrapolated for the whole data set).

The comparison of different background supression methods is shown in Table 1, where DIFF $= N_{\text{on}}^* - N_{\text{off}}^*$ is the estimate of the signal, DIFF$/N_{\text{off}}^*$ is the estimate of the signal-to-noise ratio, $N_{\text{off}}^*/N_{\text{off}}$ is the estimate of background supression by the technique used.

The Neural Net classification (multidimensional nonlinear masks method) was also applied to the detection of the Crab Nebula by another Cherenkov telescope located on La Palma (Kanarian, Hegra collaboration).

Table 2
Hegra Crab detection, 1992–1993 (Krennrich et al., 1993)

	N_{on}^*	N_{off}^*	σ	DIFF	DIFF$/N_{\text{off}}^*$	$N_{\text{off}}^*/N_{\text{off}}$
Azwidth (< 0.18)	4083	3674	4.64	409	0.11	0.05

Table 3
Hegra Crab detection, 1992, September–October

	N_{on}^{*}	N_{off}^{*}	σ	DIFF	DIFF$/N_{\mathrm{off}}^{*}$	$N_{\mathrm{off}}^{*}/N_{\mathrm{off}}$
Raw	30880	30857	0	23	0.001	
Azwidth	1333	1146	3.76	187	0.16	0.037
Width, Miss	955	784	4.1	171	0.22	0.025
Width, Miss, Length	659	492	4.99	167	0.34	0.016
Width, Miss, Length, Conc	638	460	5.37	178	0.39	0.015
Neural 4::3::1	520	342	6.08	178	0.52	0.011

The Azwidth analysis results (Krennrich et al., 1993) on this data set (150000 events) are presented in Table 2.

The new methods were applied only to part of the data (September–October 1992). In Table 3 the results of different background rejection methods as applied to this data are summarized. As one can see, the neural analysis achieves the highest signal-to-noise ratio and the smallest background contamination using 4 image parameters (Width, Length, Miss, Conc) and only 3 nodes in a single hidden layer.

4. Conclusions

There are some alternatives to the use of neural networks for background rejection in high energy physics experiments. There exist also many traditional statistical methods which are more mathematically founded, for example, Baysian statistical decisions (one can see examples of the use of this approach in high energy physics data analysis in (Chilingarian, 1989; Chilingarian and Zazyan, 1991)).

Time for training the neural net, tedious selection of network architecture, neuron output function and global learning parameters plus the dependence of results on the initial state of the network leads to the results of which optimality and reliability have to be checked with results obtained using traditional non-parametric statistical methods (Duch and Dierksen, 1994).

Many important theoretical problems of neural calculations are far from being solved. Only very few quantitative results are available. There are several practical problems to be solved:

- Selection of the learning rules for different problems;
- Investigation of the influence of the accuracy of the weights on the NN performance;
- Investigations of the role of the shape of the nonlinear output function and of the number of nodes in the hidden layer on the sensitivity of NN classifier;
- Designing fast training algorithms which minimize the true error (on a test sample) instead of minimizing the apparent error (on the training sample).

Nevertheless for solving mathematically ill-posed multidimensional nonlinear problems with ill-defined conditions (like selection of nonlinear multivariate signal domain), where common statistical methods usually fail, the use of neural techniques seem to be suitable and as one can see from the previous the section, the data analysis results can be comparable with or better than traditional techniques.

Acknowledgments

We thank the members of the Whipple and Hegra collaborations for the use of the Crab Nebula databases.

We thank Dr. M. Cawley for valuable remarks concerning new methods of data analysis, Prof. W. Hofman, Prof. F. Aharonian, Dr. A. Heusler and M. Halpaap for useful discussions.

References

Chilingarian, A.A. (1989). Statistical decisions under nonparametric a priory information. *Comp. Phys. Comm.* 54, 381–390.

Chilingarian, A.A. and M.F. Cawley (1991). Application of multivariate analysis to atmospheric Cherenkov imaging data from the Crab nebula. *Proc. 22nd Internat. Cosmic Ray Conf.*, Dublin, Vol. 1, 460–463.

Chilingarian, A.A. and H.Z. Zazyan (1990). A bootstrap method of distribution mixture proportion determination. *Pattern Recognition Lett.* 11, 781–785.

Chilingarian, A.A. and H.Z. Zazyan (1991). On the possibility of investigation of the mass composition and energy spectra of PCR in the energy range 1–10 eV using EAS data. *Il Nuovo Cimento* 14C (6), 555–568.

D0 Collaboration, S. Abachi et al. (1994). The D0 detector. *NIM* A338, 185.

Denby, B. (1992). Tutorial on neural network applications in high energy physics: a 1992 perspective. Fermilab-Conf-92/121-E.

Duch, W. and G.H.F. Dierksen (1994). Neural networks as tools to solve problems in physics and chemistry. *Comp. Phys. Comm.* 82, 91–103.

Flugge, G. (1991). Physics at the large hadron collider. *Proc. 1991 CERN School of Computing*, 23–50.

Fogelman Soulie, F. (1992). Neural networks for pattern recognition: introduction and comparison to other techniques. In: D. Perret-Gallix, Ed., *Proc. 2nd Internat. Workshop on Soft. Enginer.*, 277–286.

Hillas, M. (1985). Cherenkov light images of EAS produced by primary gamma rays and by nuclei. *Proc. 19th Internat. Cosmic Ray Conf.*, USA, Vol. 3, 445–448.

Krennrich, F., R. Mirzoyan, et. al. (1993). Observation of VHE gamma emission from the Crab nebula with the prototype of the Hegra air Cherenkov telescope array. *Proc. 23rd Internat. Cosmic Ray Conf.*, Calgary, Vol. 1, 251–254.

Lang, M.G., C.W. Akerlof, M.F. Cawley, et. al. (1991). TeV observation of the Crab nebula and other plerions in the epoch 1988–91. *Proc. 22nd Internat. Cosmic Ray Conf.*, Dublin, 204–207.

Mapelli, L. (1991). Collisions at future supercolliders: the first 10 microseconds. *Proc. 1991 CERN School of Computing*, 79–112.

Peterson, C. and T. Gvaldsson (1991). An introduction to artificial neural networks. *Proc. 1991 CERN School of Computing*, 113–170.

Punch, M., C.W. Akerlof, M.F. Cawley, et. al. (1991). Supercuts: an improved method of selecting gamma-rays. *Proc. 22nd Internat. Cosmic Ray Conf.*, Dublin, Vol. 1, 464–467.

Ruck, D.W., K.S. Rogers, et al. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Trans. Neural Networks* 1, 296.

Vacanti, G., M.F. Cawley, et. al. (1991). Gamma-ray observations of the Crab nebula at TeV energies. *Astroph. J.* 377, 467–475.

Zhang, S.N. and D. Ramsden (1990). Statistical data analysis for g-ray astronomy. *Exp. Astronomy* 1, 145–158.