Neural networks as classifiers in subatomic physics

CARSTEN PETERSON

Department of Theoretical Physics
University of Lund
Sölvegatan 14A
S-22362 Lund, Sweden

Introduction

When analyzing experimental data it is standard procedure to make various cuts in observed kinematical variables in order to single out desired features. A specific choice of cuts corresponds to a particular set of feature functions $o_i = F_i(x_k)$ in terms of the kinematical variables x_k. This procedure is often not very systematic and quite tedious. Ideally one would like to have an automated optimal choice of the functions F_i. This is exactly what feature recognition artificial neural networks (ANN) aim at. For feed-forward ANN the following form of F_i is often chosen

$$F_i(x_k) = g\left[\sum_j w_{ij} g\left(\sum_k w_{jk} x_k\right)\right] \qquad (1)$$

where w_{ij} and w_{jk} are the parameters to be fitted to the distributions and

$$g(x) = 0.5\,[1 + \tanh(x)] \qquad (2)$$

Eq. 1 corresponds to the feed-forward architecture of Figure 1. The bottom layer (input) corresponds to the measured kinematical variables x_k and the top layer to the features o_i. The mission of the so-called hidden layer is to build up an internal representation of the observed data. Eq. 1 and Figure 1 are easily generalized to more than one hidden layer. Each unit or neuron has the threshold behaviour given by $g(x)$. Fitting to a given data set (or learning) takes place with gradient descent on, for example, a summed square error

$$E = \frac{1}{2}\sum_i (o_i - t_i)^2 \qquad (3)$$

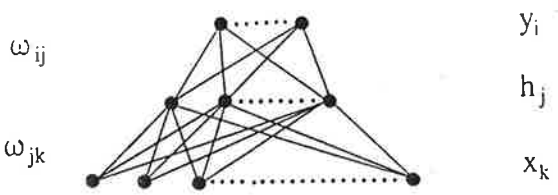

Figure 1. *A one hidden layer feed-forward neural network architecture.*

with respect to the weights w_{ij} and w_{jk}, where t_i are the desired feature values. In this process, which is called *back-propagation (BP)* (1), the *training patterns* are presented over and over again with successive adjustments of the weights. Once this iterative learning has reached an acceptable level in terms of a low error E the weights are frozen and the ANN is ready to be used on patterns it has never seen before. The capability of the network to correctly characterize these test patterns is called generalization performance.

This back-propagation procedure assumes the knowledge about what features (o_i) are relevant from the outset and separates the data accordingly. There is also an alternative approach, self-organization, where the network organizes the data into features without any external teacher (no output units) (2). The underlying architecture consists of an input layer (x_k) and a layer of feature nodes denoted h_j (see Figure 2), where

$$h_j = g\left(\sum_k w_{jk} x_k\right) = g\left(\vec{w}_j \vec{x}\right) \qquad (4)$$

For all patterns presented the weights are updated such that the angles between $\vec{w}_j$ and $\vec{x}$ are minimized. Also, topological ordering between the feature nodes h_j are introduced with a "mexican-hat" potential, such that neighbouring nodes in a plane react to similar features. Such a system has no "teacher" like the feed-forward BP network above where answers are compared with correct values t_i. The results in this approach are extremely easy to analyze; the weight vectors $\vec{w}_j$ for the different feature nodes point in the direction of typical data in $\vec{x}$-space.

The ANN approach has turned out to be very profitable for a wide range of pattern recognition problems ranging from identification of handwritten numerals (3), transformation of written text to speech (4) to quark identification in particle physics (5,6,7,8). The latter domain of applications is the focus of this paper. Most applications described here are done

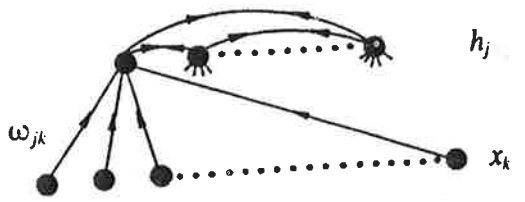

Figure 2. *A one-layer self-organizing network. Lateral interactions between the feature nodes correspond to the "mexican-hat" potential.*

with the BP algorithm since it seems to give the best performance in terms of classification percentages.

Identifying the Origin of Hadronic Jet

Background

In high energy lepton-lepton, lepton-hadron, hadron-hadron and nucleus-nucleus collisions, quarks and gluons are produced. These basic quanta of Quantum Chomodynamics (QCD) can never be observed due to the confinement mechanism. They fragment into jets of hadrons, which are observed in the laboratory. Currently, this fragmentation process can not be calculated from first principles, but theoretically plausible Monte Carlo Markov models (9,10,11) have been developed that reproduce the data extremely well. The main problem here is given a pattern of hadrons in terms of their kinematical variables, unfold their origin by quark species or gluon identity. In e^+e^- reactions, the kinematical information is typically given in terms of energy and momenta, whereas in hadron-induced reactions calorimeter detectors pick up transverse energy deposits (E_T) in different cells. The cell positions are given by azimuth angle ϕ and pseudorapidity η (the latter is to a good approximation the velocity of the produced hadrons in the beam direction).

Quark-gluon separation

Being able to distinguish whether a jet of hadrons originates from a quark or a gluon is important from many perspectives. It can shed light on the hadronization mechanism. For example, experimental studies of the so-called string effect require identification of the gluon jet. In addition, a fairly precise identification of the gluon jet is required for establishing the existence of the 3-gluon coupling in e^+e^- annihilation reactions. In refs. 5, 6 and 7 the BP algorithm was used to do the separation both in e^+e^- reactions and in hadron-induced large p_T processes.

e^+e^- reactions (5, 6). In order not to reveal too much about the MC model dependent low momentum part of the jet, four-momenta $(\vec{p}_k, E_k)$ of the four leading particles in the jet were used as inputs to the network. One output node representing the quark/gluon identity was used, as well as 10 hidden units. Training and test sets were generated at 2 different energies (92 GeV and 29 GeV) with 3 different MC generators; JETSET (9), ARIADNE (11) and HERWIG (10). The inputs were either single jets defined by the LUCLUS clustering algorithm in JETSET (9) or entire 3-jet events. After training was completed the network was tested with a middle point succes criteria where > 0.5 for the output node is interpreted as a gluon jet and < 0.5 as a quark jet.

On the average the network was able to correctly classify 85% of the test set jets. The MC model independence of the results was demonstrated by training on MC data generated by one model and tested on MC data from another. Almost no deterioration in performance was observed. Runs where detector acceptance effects were included showed only $O(2\%)$ degradation.

The QCD matrix element suppresses gluon jet production as $1/E_{gluon}$. A fair part of the ANN discrimination originates from this property. In order to factor out this matrix element dependence from the intrinsic differences between quark and gluon jets, different networks should be trained with quark and gluon jets in different energy intervals, combining the answers with the appropriate matrix element predictions. With such a procedure, the classification power increases to 92% (12).

Large p_T processes (8). In this case, the network gets no lead from QCD matrix element information, since the kinamatics of the incoming quarks are unknown. A lower classification performance can be expected. Also, in hadron-hadron collisions the momenta and energies of the produced hadrons are available in a "raw" form in terms of towers in a calorimeter representing the transverse energies E_T, as mentioned previously. In ref. 7, a set of $p\bar{p}$ events at 630 GeV was generated with the PYTHIA MC (13). The transverse energy of the fragmentation products was mapped into a calorimeter with a granularity of $\Delta\eta = 0.20$ in pseudorapidity (≈longitudinal velocity) and $\Delta\varphi = 0.26$ in azimuth. This calorimeter had a complete coverage in φ and extends out to $|\eta| \leq 2$. The set-up corresponds to the UA2 calorimeter at CERN. Jet transverse energies were collected in cones of radius 0.8 in η,φ space. The calorimeter information was presented to the network as follows: Take the E_T of the leading cell in the 7×7 matrix and assign it to the first node x_1. Assign the η and φ coordinates relative to the center of the jet to x_2 and x_3 respectively. Then take the second leading cell and assign its E_T, η and φ to x_4, x_5 and x_6 and so on for the first 15 cells. This corresponds to 45 input nodes. The reason for choosing this representation of the input data rather than the 7×7 cells directly is that in this way, invariances of the data are more efficiently incorporated.

After training, the network correctly classifies 70 – 72% of the jets using the 0.5 criteria. Rather than having this success rate, the cut of the output node should be varied and a value corresponding to an optimal efficiency and signal-to-background ratio should be chosen. Figure 3 shows the signal-to-background increase as a function of the signal efficiency. In ref. 7 a similar encoding was successfully used for separating jets stemming from the intermediate vector boseon W from those originating from QCD collision processes. Such a network can reduce the QCD background to $W/Z\rightarrow$ jets by factors 20 – 30.

Heavy quark jet tagging

So-called heavy quarks (c and b) are produced at high energies. These are unstable and decay, weakly emitting leptons. The conventional way of identifying heavy quarks in e^+e^- reactions is either through leptonic tagging or secondary

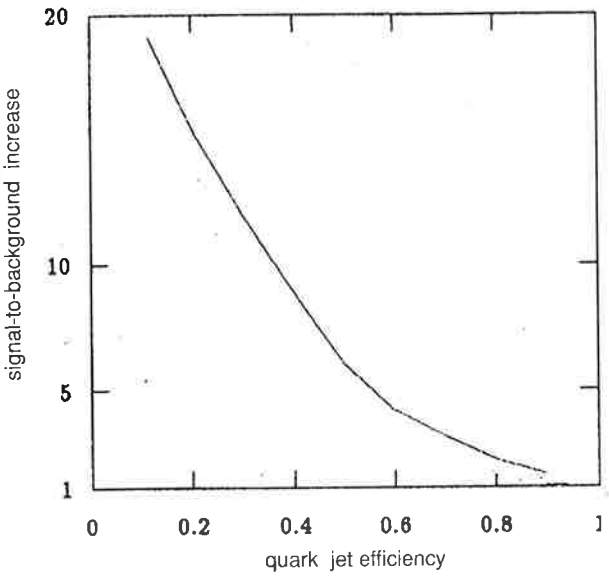

Figure 3. *The signal-to-backround ratio increase versus the efficiency for a neural network quark trigger.*

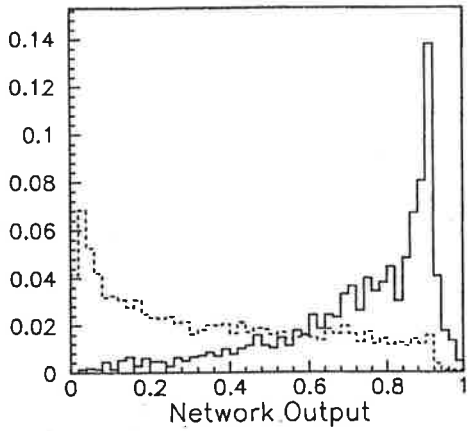

Figure 4. *Output distributions for b (full line) and non-b (dashed line) hadronic jets (from ref. 14).*

vertex with efficiency/purity levels of approximately 5%/95% and 25%/80% respectively. We have designed a network that identifies heavy quarks (6), which is entirely based upon hadronic information. Total jet energy and momentum along with the energy and direction for the 6 leading particles in the jet are used as input variables, giving a total of 20 input units. Again, one layer of 10 hidden units is used with a single output neuron (b=1, non-b=0). As in the large p_T application, the option of varying the cut on the output node can be used to select a desirable efficiency versus purity. The network is then able to produce efficiency/purity numbers comparable with what is expected from vertex detectors. This remarkable result implies that a general purpose hadronic detector could also prove very efficient for heavy quark tagging. Even better efficiency/purity ratios were obtained in subsequent work (14) by preprocessing the kinematic variables in terms of different shape variables. In Figure 4 the distribution of events for the output node is shown from ref. 14. It is clear from this figure how choosing the output threshold governs the efficiency/purity ratio.

Studying heavy quark tagging in a self-organizing network can prove very illuminating. In ref. 8, such a network, consisting of a plane of 7×7 feature nodes, was used to disentangle b-, c- and light (uds) quarks. The resulting distributions of the feature nodes are shown in Figure 5 while Figure 6 reveals the mapping in terms of dominating quarks.

Not surprisingly, the two extremes in terms of quark masses occupy two distinct areas separated by the c-quarks. It is very interesting to inspect the corresponding weight vectors $\vec{w}_j$. For example, the nodes in the lower right corner have $\vec{w}_j$'s with an even distribution of momenta (see Figure 7), which is exactly what one expects for b-quarks since they decay more isotropically.

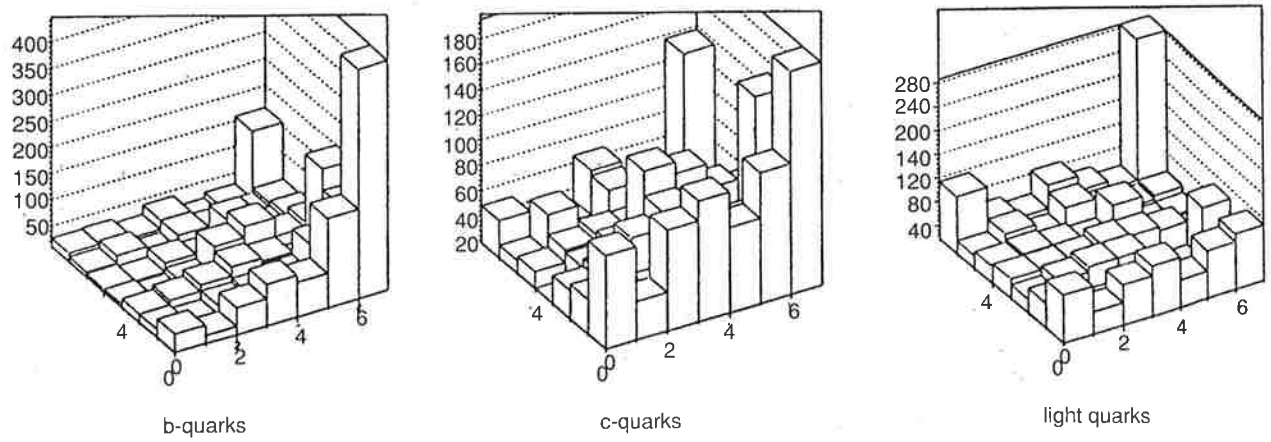

Figure 5. *Distribution of b-quarks (a), c-quarks (b) and uds-quarks (c) over the self-organized 7×7 feature nodes.*

Mass reconstruction

The problem of computing invariant masses of expected decay products often appears when hunting for new particles or resonances. For example, in the case of the intermediate vector boson W (produced in $\bar{p}p$ collisions in its hadronic decay channel, $W \rightarrow \bar{q}q \rightarrow$ hadrons, M_W is reconstructed relativistically from the momenta and assumed masses of the produced hadrons. The problem here is that the q and q jets are not the only hadrons in the collision—there are also remnants from projectile hadrons. An additional complication is that the q and/or $\bar{q}$ jets can give rise to additional jets through bremsstrahlung, so identifying the appropriate jets is crucial for a good reconstruction of the mass. The "standard" procedure (15) for doing this is by sweeping through the calorimeter with a "window" of a certain (φ,η) size. The two windows with the largest total E_T are selected as containing the two jets and the hadrons in these jets are used to compute M_W.

The neural network approach to this problem is as follows (16). Since this is not a classification task, a linear rather than nonlinear (eq. (2)) output node is used for the answer (M_W). The 30 largest towers from the calorimeter are used as inputs (cf. quark/gluon separation above) together with the total E_T. A network with two hidden layers with 36 and 15 hidden units respectively is trained with the BP algorithm. As a training set MC generated data with a flat distribution of M_W in the range [50,150] GeV is used. When tested on "real data" in terms of MC generated realistic W-masses, the ANN approach produces a distribution which is more narrow and symmetric than the one using conventional methods (see Figure 8). The main reason why the ANN method does better than the conventional method is that it captures gluon bremsstrahlung well.

Feedback networks—track finding

The networks discussed here have been based on feed-forward architectures with classification as the main application. In feedback architectures, information flows both ways through the weights and the final state is given by a fixed-point solution of the neurons v_i that minimizes an energy function.

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} v_i v_j \qquad (5)$$

Such networks have been successfully applied to difficult optimization problems (17). The weights w_{ij} in this case are not adaptive parameters —they are fixed constants set by the optimization problem in such a way that an optimal solution minimizes eq. 5. The neurons v_i encode different solution possibilities. Reconstructing tracks out of signals is a common optimization problem in subnuclear physics, where the feedback ANN approach and variants thereof have shown great promise (18-21). Since feedback networks are natural to

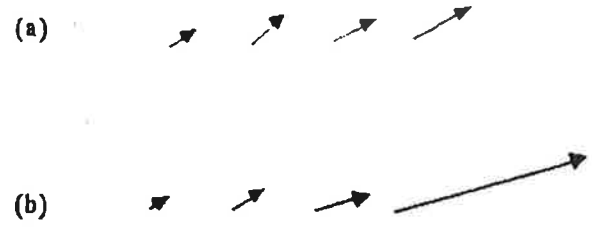

Figure 7. The weight vectors corresponding to the 4 leading hadrons for b-sensitive (a) and uds-sensitive (a) units in the self-organizing network. The pT component has been multiplied by a factor of 5 relative to the px component.

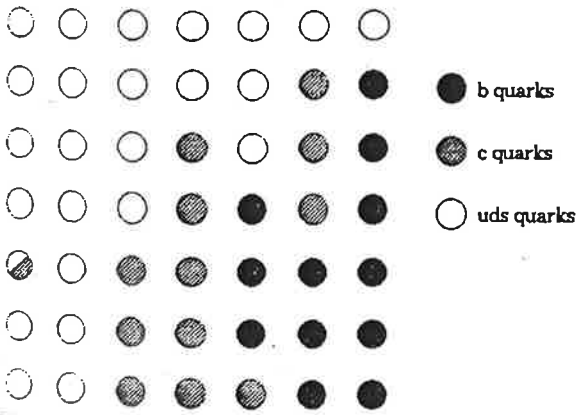

Figure 6. The resulting map for uds- c- and b-quarks. The shading indicates the dominant flavor for the units. The units are numbered as in Figure 5.

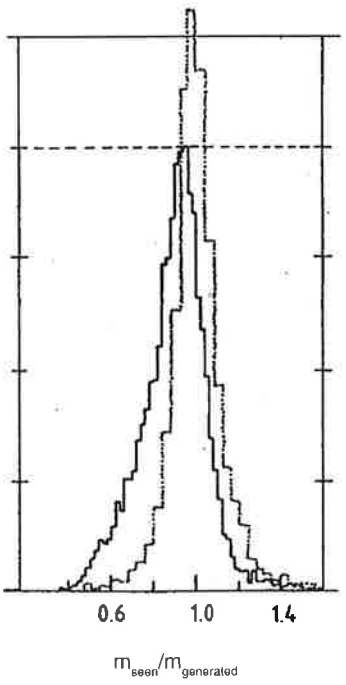

Figure 8. The reconstructed mass (M_W) divided by the true mass (M^0_W) using the ANN method (dotted line) and the conventional "window" method (full line).

implement in VLSI, a very good candidate algorithm for O(*ns*) real-time track finding is thus available.

Summary and outlook

The neural network method is very efficient for extracting features in hadronic data. World record performance can be obtained for quark/gluon separation. With respect to heavy quark tagging, the results are in parity with those expected from a vertex detector. A similar network is also able to reduce the QCD background to $W/Z \rightarrow$ jets by a factor of 20–30.

All results reported here (5,6,7,8) were obtained by using a JETNET F77 software package (22). It should be emphasized that the feed-forward ANN approach involves nothing more than fitting data with layers of sigmoids (cf. eq. 2). For high-dimensional problems this is more effecient than the more common approach of using Gaussian expansions. In addition, sigmoidal amplifiers are easily available VLSI devices, which facilitates hardware implementations.

The neural network approach is in general very noise- and damage-resistant, making it suitable for high energy experiments where various parts of a detector may malfunction. With its inherent concurrency and simple structure, fast execution custom-made hardware could also be an asset for real-time triggering. In addition to feature recognition, ANN will very likely play an important role in track finding, since feedback networks easily lend themselves to real-time hardware implementations.

—received 29 July 1991

References

1. D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning Internal Representations by Error Propagation," in D. E. Rumelhart and J. L. McClelland (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 1), MIT Press (1986).
2. T. Kohonen, "Self Organized Formation of Topologically Correct Feature Maps," Biolological Cybernetics 43, 59 (1982); T. Kohonen, Self-organization and Associative Memory, third edition, Springer-Verlag, Berlin, Heidelberg (1990).
3. Y. LeCun et. al., "Backpropagation Applied to Hadwritten Zip Code Recognition," Neural Computation 1, 541 (1989).
4. T.J. Sejnowski and C.R. Rosenberg, " Parallel Networks that Learn to Pronounce English Text," Complex Systems 1, 145 (1987).
5. L. Lönnblad, C. Peterson and T. Rögnvaldsson, "Finding Gluon Jets with a Neural trigger", Physical Review Letters 65, 1321 (1990).
6. L. Lönnblad, C. Peterson and T. Rögnvaldsson, "Using Neural Networks to Identify Jets," Nuclear Physics B349, 675 (1991).
7. P. Bhat, L. Lönnblad, K.Meier and K.Sugano, "Using Neural Networks to Identify Jets in Hadron-Hadron Collisions," LU TP 90-13 (to appear in Proc. of the 1990 DPF Summer Study on High Energy Physics Research Directions for the Decade, Colorado, 1990).
8. L. Lönnblad, C. Peterson, H. Pi and T. Røgnvaldsson, "Self-organizing Networks for Extracting Jet Features," LU TP 91-4 (to appear in Computer Physics Communications).
9. T. Sjöstrand, JETSET 7.2 program and manual. See B. Bambah et al., QCD Generators for LEP, CERN-TH.5466/89.
10. G. Marchesini and B. R.Webber, Nuclear Physics B310 (1988) 461; I. G. Knowles, Nuclear Physics B310, 571 (1988).
11. L. Lönnblad, "ARIADNE-3, A Monte Carlo for QCD Cascades in the Colour Dipole Formulation," Lund preprint LU TP 89-10.
12. I. Scabai, F. Czakó and Z. Fodor, "Quark and Gluon Jet Separation using Neural Networks," ITP Budapest Report 477 (1990).
13. H-U. Bengtsson, T. Sjöstrand, Computer Physics Communications 46, 43 (1987).
14. L. Bellantoni, J.Conway, J. Jacobsen, Y.B. Pan and S.L. Wu, "Using Neural Networks with Jet Shapes to Identify *b* Jets in e^+e^- Interactions", CERN-PPE/91-90 (submitted to Nuclear Instruments and Methods A).
15. J. Alitti et al., "Measurements of the Transverse Momentum Distributions of *W* and *Z* Bosons at the CERN $\bar{p}p$ Collider", Zeitschrift für Physik C47, 523 (1990).
16. L. Lönnblad, C. Peterson and T. Rögnvaldsson, "Mass Reconstruction with a Neural Network," LU TP 91-25 (to appear in Physics Letters B).
17. C. Peterson, "Parallel Distributed Approaches to Combinatorial Optimization," Neural Computation 2, 261 (1990).
18. Denby, B. "Neural Networks and Cellular Automata in Experimental High Energy Physics," Computer Physics Communications 49, 429 (1988).
19. C. Peterson, "Track Finding with Neural Networks," Nuclear Instruments and Methods A279, 537 (1989).
20. M. Gyulassy and H. Harlander "Elastic Tracking and Neural Network Algorithms for Complex Pattern Recognition," Computer Physics Communications 49, 429, 1988.
21. A. Yuille, K. Honda and C. Peterson, "Deformable Templates for Particle Tracking," Proceedings of 1991 IEEE INNS IInternational Joint Conference on Neural Networks, Seattle (IEEE 91CH3049-4).
22. L. Lönnblad, C. Peterson and T. Rögnvaldsson, "Pattern Recognition in High Energy Physics with Artificial Neural Networks—JETNET 2.0," Lund Preprint LU TP 91-18 (to appear in Computer Physics Communications) [Program and manual available via email request (carsten@thep.lu.se)].