Disclaimer

This note has not been internally reviewed by the DØ Collaboration. Results or plots contained in this note were only intended for internal documentation by the authors of the note and they are not approved as scientific results by either the authors or the DØ Collaboration. All approved scientific results of the DØ Collaboration have been published as internally reviewed Conference Notes or in peer reviewed journals.

THE RANDOM GRID SEARCH: A SIMPLE WAY TO FIND OPTIMAL CUTS *

N. AMOS, C. STEWART†
Department of Physics, University of Michigan,
Ann Arbor, MI 48109, USA

P. BHAT
Fermi National Accelerator Laboratory
PO Box 500, Batavia, IL 60510, USA

C. CRETSINGER, E. WON
Department of Physics, University of Rochester,
Rochester, NY 14627, USA

W. DHARMARATNA‡, H.B. PROSPER
Department of Physics, Florida State University
Tallahassee, FL 32306, USA

Abstract

With the availability of cheap powerful computers multivariate methods of analysis that hitherto were impractical are now feasible. We describe a method, which is being used by the DØ collaboration, to find optimal cuts for any given n-tuple of event variables. The method, which we have dubbed the *random grid search*, is an efficient variant of the search for cuts on a regular grid. The results of this method compare favorably with those from a feed-forward neural network. This is illustrated using the extraction of a top quark signal from multi-jet data, as an example.

---

*Talk given at the 1995 *Computing in High Energy Physics Conference* (CHEP '95), Rio de Janeiro, Brazil.

†Now at BioLogic

‡Permanent address: Dept. of Physics, University of Ruhuna, Matara, Sri Lanka

I. INTRODUCTION

A basic task in high energy physics research is to classify an event as either signal or background. An event is described by an n-tuple $x = (x_1, \ldots, x_n)$ of variables, which variables are chosen—using physical intuition—to capture the essential features of the events. The n-tuples populate an n-dimensional space called, appropriately, *feature space.* The mathematical task then is to divide the feature space into regions that best isolate the signal events from the background. Intuition suggests that the best way to find the boundaries between these regions is to minimize the probability to mis-classify signal and background events. One can regard the various event classification methods as simply different ways to approach this ideal. Here we compare two methods: classification using a a feed-forward neural network and classification using a simple variant of the grid search.

II. EVENT CLASSIFICATION

A. In Principle

The particle reactions—that is, events—studied today are seldom of the sort that can be classified unambiguously as either signal or background. We must resort to probabilistic methods of classification that, in princple, require knowledge of the probability density functions $f(x|s)$ and $f(x|b)$, respectively, of the signal and background n-tuples. In addition, we may need the prior probabilities $p(s)$ and $p(b)$, respectively, for the signal and background. The quantity $p(s)/p(b)$ is the signal to background ratio before event classification. We note that $p(s) + p(b) = 1$, $\int dx f(x|s) = 1$ and $\int dx f(x|b) = 1$.

The boundaries that minimize the probability to mis-classify are obtained as solutions of the equation

$$r(x) = \frac{f(x|s)p(s)}{f(x|b)p(b)} = c, \tag{1}$$

where c is a constant. The quantity $r(x)$ is called the *Bayes discriminant function.* It bears a simple relationship to Bayes' theorem:

$$p(s|x) = \frac{r}{1+r}. \tag{2}$$

The probability $p(s|x)$ is precisely that required to classify events: it is the probability that an event is of the signal class given that it is characterized by the n-tuple x. Notice the logical distinction between $p(s|x)$, the *Bayesian posterior probability,* and the probability $p(x|s) \equiv f(x|s)dx$.

The equation $r(x) =$ constant, or equivalently $p(s|x) =$ constant defines hypersurfaces between signal and background regions that separate these regions optimally. These hypersurfaces are referred to as *decision boundaries.* Observe that we can always absorb the prior probabilities into the constant in equation 1; so we don't really need these priors, in which case the Bayes discriminant function $r(x)$ reduces to the well-known likelihood ratio, and $p(s|x)$ simplifies to

$$p(s|x) = \frac{f(x|s)}{f(x|s) + f(x|b)}. \tag{3}$$

B. In Practice

It is generally impossible to write down an analytic form for the n-dimensional densities $f(x|s)$ and $f(x|b)$ and so it might appear that our project to compute $r(x)$ is doomed to failure. Happily this is not the case for the following reason. Under suitable circumstances, that we shall not comment upon here, the output of a feed-forward neural network provides a direct approximation to the probability $p(s|x)$ [1]. But, while it is true that this result and others [2] provide a sound mathematical underpinning for neural networks, this is somewhat offset in practice by the difficulty of assessing the quality of the approximation. That observation, coupled with its (unwarranted [3]) "black box" stigma, fuels the search for better ways of classifying events with conventional methods that, nonetheless, are multivariate in spirit.

III. THE RANDOM GRID SEARCH

Conventional analyses in high energy physics separate signal from background by applying a set of cuts $x_1 > z_1$, $x_2 > z_2, \ldots$. We shall refer to $(z_1, \ldots, z_n)$ as a cut-point. The cuts are usually arrived at by a laborious process of trial and error moderated by common sense. Unfortunately, there is no guarantee that this procedure will lead to optimal cuts. A better way to obtain optimal cuts is to perform a systematic search for them over a grid of points.

A search over a regular grid, however, is inefficient: a lot of time can be spent scanning regions of feature space that have few signal or background points while spending the same amount of time scanning regions that are dense in points. Moreover, the number of grid points grows rapidly with bin count and dimensionality (like the number of bins raised to the power of the number of dimensions). It would be more efficient to put most of the computing cycles where there are most points.

The best way to do that is to use the actual distribution of points $x = (x_1, \ldots, x_n)$ as a set of cut-points. The points could, for example, have been generated by a Monte Carlo simulation of the events. The set of cut-points forms a grid with random spacing between lines, that is, a *random grid.* In the example described below we used the distribution of the signal points as the set of cut-points.

IV. AN EXAMPLE: $T\bar{T} \rightarrow \geq$ 6JETS

An important objective of the DØ collaboration is to study the decay modes of the recently discovered top quark [4]. The most challenging mode is that in which the top quark decays into a b quark and a W boson with the W bosons decaying hadronically. In this mode the signal to be extracted is tiny compared with the QCD multi-jet background. It is, therefore, of the utmost importance to find optimal cuts. For this the random grid search can be used to good effect as we now illustrate.

We consider a 3-dimensional feature space containing the n-tuples $x = (C, A, N)$ where $C = \sum |E_{Tj}| / \sum E_j$ is the centrality, $A = 3/2 \times$ smallest eigenvalue of $\sum p^a p^b / \sum p^2$ is the aplanarity and $N = \sum |e_{Tj}| N_j / \sum |e_{Tj}|$ we refer to simply as the "jet count" [5]. The quantities E_T, E and p are, respectively, the jet transverse energy, the jet energy and the jet momentum; N_j is the number of jets above the jth transverse energy threshold e_{Tj}. The sums are over all jets in the event. These variables have been found to provide a useful degree of separation between signal and background. For example, figure 1 shows the degree of separation between the signal and background in the variables N and A. The signal is simulated $t\bar{t}$ decays to jets, assuming a top quark mass of 180 GeV/c^2; the background is from multi-jet QCD data.

The grid search was done with 5000 signal events as the supplier of cut-points. With 5000 background events and another 5000 signal events we counted the number of signal and background events S_j and B_j, respectively, that passed the cuts specified by the cut-point $z_j = (C, A, N)_j$. This was repeated for every cut-point. These data can be usefully displayed in the unit square with the signal fraction, $S_j/5000$, on the vertical axis and the background fraction, $B_j/5000$, running horizontally. The entire calculation took about 40 CPU seconds on a DEC Alpha 3000 Model 600 workstation.

A similar calculation was performed on the (1-dimensional) output of a feed-forward neural network having 3 input nodes, 5 nodes in a single hidden layer and one output node. That is, each output value, which approximates the posterior probability $p(s|C, A, N)$, was taken as a cut-point and the corresponding signal and background fractions were calculated. The network was trained with 2000 iterations, using the JETNET program (V3.0) from the University of Lund, and using 5000 events divided equally between signal and background. This required about 300 CPU seconds on the same model of workstation. The results of the random grid search and neural network are compared in figure 2.

The outer envelope of the points pertaining to the random grid search (the open circles in figure 2) defines the set of optimal cuts for the given n-tuple of variables. The network curve, as expected, is higher than this envelope. But, given its simplicity the random grid search does remarkably well. This is especially true for rejection factors against background of 30 or higher.

This example illustrates that with a modest amount of work it is possible to put to good use the considerable computing power that is now at our disposal. We have seen that a simple grid search, modified as indicated, requires no more CPU time to execute than that required to train a feed-forward neural network. Moreover, the optimal cuts obtained are applied directly to the orignal physics–motivated variables and as such are immediately intelligible.

ACKNOWLEDGMENTS

We thank our colleagues at DØ for many useful discussions. This research is supported in part by the U.S. Department of Energy.

REFERENCES

[1] D.W. Ruck *et al.*, The multilayer perceptron as an approximation to a Bayes optimal discriminant function, *IEEE Trans. Neural Networks* **1 (4)** (1990) 296; E.A. Wan, Neural network classification: a Bayesian interpretation, *IEEE Trans. Neural Networks* **1 (4)** (1990) 303.

[2] E.K. Blum and L.K. Li, Approximation theory and feedforward networks, *Neural Networks*, **4** (1991) 511.

[3] See, for example, J. Linnemann, these proceedings.

[4] DØ Collab., S. Abachi *et al.*, Observation of the Top Quark, *Phys. Rev. Lett.* **74** (1995) 2632. CDF Collab., A. Abbe *et al.*, Observation of Top Quark Production in $\bar{p}p$ Collisions with the Collider Detector at Fermilab, *Phys. Rev. Lett.* **74** (1995) 2626.

[5] We thank Fyador Tkachov for pleasant and instructive discussions that inspired one of us (C.S.) to invent this variable. F. Tkachov, *Phys. Rev. Lett.* **73** (1994) 2405.

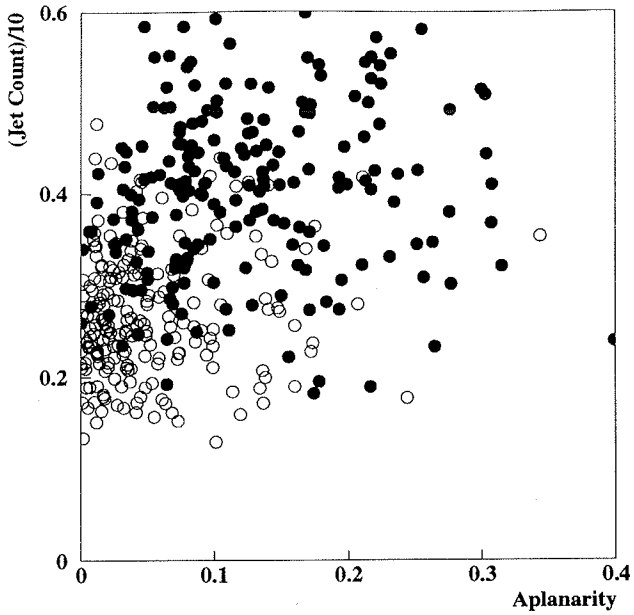

FIG. 1. (Jet count)/10 vs. aplanarity. The open circles are background points and the closed circles pertain to signal.

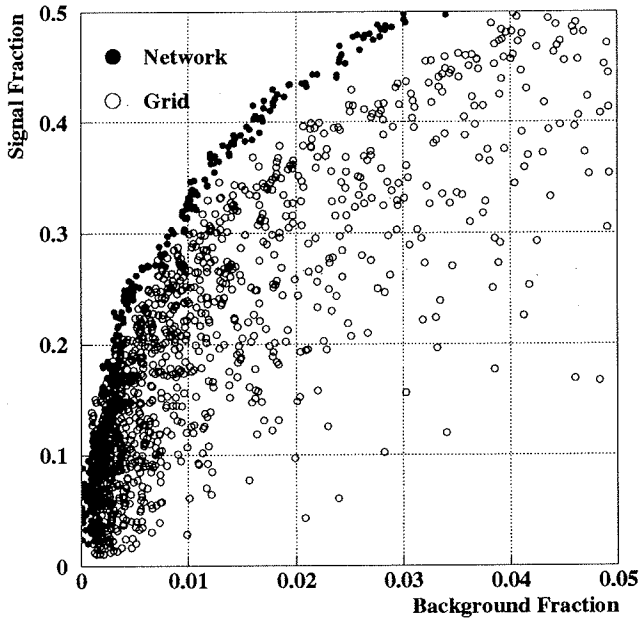

FIG. 2. Signal fraction vs. background fraction for the grid (open circles) compared with the results from a network (closed circles). Note the large spread in the points computed with the grid. It is a warning that the phase space for finding poor cuts is much larger than that for finding optimal ones!