BENEFITS OF MINIMIZING THE NUMBER OF DISCRIMINATORS USED IN A MULTIVARIATE ANALYSIS[*]

S. Towers
State University of New York at Stony Brook

Abstract

As particle physics experiments grow more complicated with each passing decade, so too do the analyses of data collected by these experiments. Multivariate analyses involving dozens of variables are not uncommon in this field. This note describes how the use of many variables in a multivariate analysis can actually degrade the ability to distinguish signal from background, rather than improve it. A method which can aid in reducing the number of variables to an optimal set of discriminators is also described.

1 Introduction

High energy physicists today are usually occupied with the task of trying to extract small signals (representing the interesting physics) out of an almost overwhelming sea of background events. To achieve this difficult goal, the particle physics community has become conversant in recent years with a number of sophisticated multivariate techniques. Studies of top quark production are examples of complicated analyses that profit from the application of multivariate methods[1].

However, a problem with some of these multivariate techniques, and particularly endemic with neural networks, is that it is usually quite easy to include many variables in an analysis. Previous to the advent of neural network techniques in particle physics, taking into account intricate non-linear correlations between more than two or three variables was prohibitively complicated, and thus much effort was normally expended trying to find the best possible set of two or three discriminators to use in an analysis. Now, however, it is not uncommon to find neural network analyses that involve literally dozens of variables. The pervading philosophy seems to be that having a lot of variables can't do any harm, and may possibly yield extra discrimination between signal and background from the complex intercorrelations assumed to exist between the variables.

Obviously, using many variables in an analysis makes it quite difficult to determine if the signal+background model is accurately describing the data in the multidimensional parameter space. In addition to this problem, however, adding too many weakly discriminating variables to a multivariate analysis will actually degrade, rather than enhance, the ability to distinguish between signal and background. This is because any added variable may (or may not) add discrimination between signal and background, but will **always** add statistical noise. As we will see in a moment, this degradation of the ability to distinguish between signal and background will occur no matter which multivariate analysis technique is being used.

A simple example of this phenomenon is as follows: a sample of 'signal' events is generated using a five-dimensional Gaussian probability density function (PDF). A sample of 'background' events is also generated with a five-dimensional Gaussian PDF, that is identical in every way to the signal PDF, except that the mean in one dimension is shifted by one standard deviation from the signal mean in that dimension. This particular dimension will be called the 'useful' dimension. The other four dimensions, for which the signal and background PDF's are identical, will be referred to as the 'useless' dimensions.

Figure 1 shows the background efficiency versus signal efficiency when only the useful dimension is used to discriminate between signal and background. The background efficiency versus signal efficiency curve is a measure of the discrimination power of a multivartiate analysis; the smaller the area

[*] A colour version of this paper is available at `http://www-d0.fnal.gov/~smjt/durham/reduc.ps`

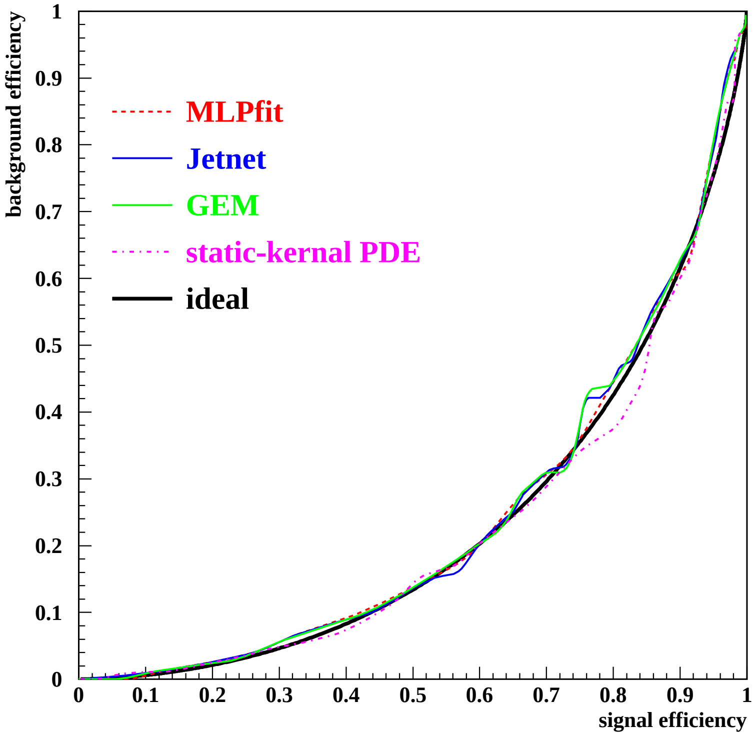

Fig. 1: Background efficiency versus signal efficiency as obtained by four different multivariate techniques under the hypothesis that the signal and background are both unit-width one-dimensional Gaussians separated by one unit.

under the curve, the better the general discrimination power. In this case, the discrimination is approximately the same for four different multivariate techniques [2, 5], and all approach the ideal, as indicated by the heavy black line.

Figure 2 shows the background efficiency versus signal efficiency when the four useless dimensions are added to the analysis. The discrimination power of all the multivariate techniques is significantly degraded. To avoid this effect, a technique is needed to reduce the number of variables used in the analysis to a subset that provides optimal discrimination between signal and background.

2 Reducing the Number of Variables

Here I will present two methods to reduce the number of variables. The first is simple-minded, and can quickly sort through a list of variables to find many of the best discriminators. The second is slightly more complicated, but is more stable when the size of the training samples is changed.[1]

2.1 The Simple Algorithm

The user begins the algorithm with a set of variables (possibly numerous) that are believed to perhaps afford some distinction between signal and background. For each variable, the user performs a univariate analysis (using the multivariate analysis method of their choice), determining the $S/\sqrt{S+B}$ discrimination statistic. The user then chooses the one variable that appears to afford the best discrimination between signal and background, and also adequately models the behaviour of the data. This variable forms the nucleus of “the accepted set of variables”.

Now, an iterative process begins:

[1] Other, more complicated, methods than these are available, such as genetic algorithms. The methods described here have the advantage that they are relatively easy to implement, and also easy for the average person to understand.

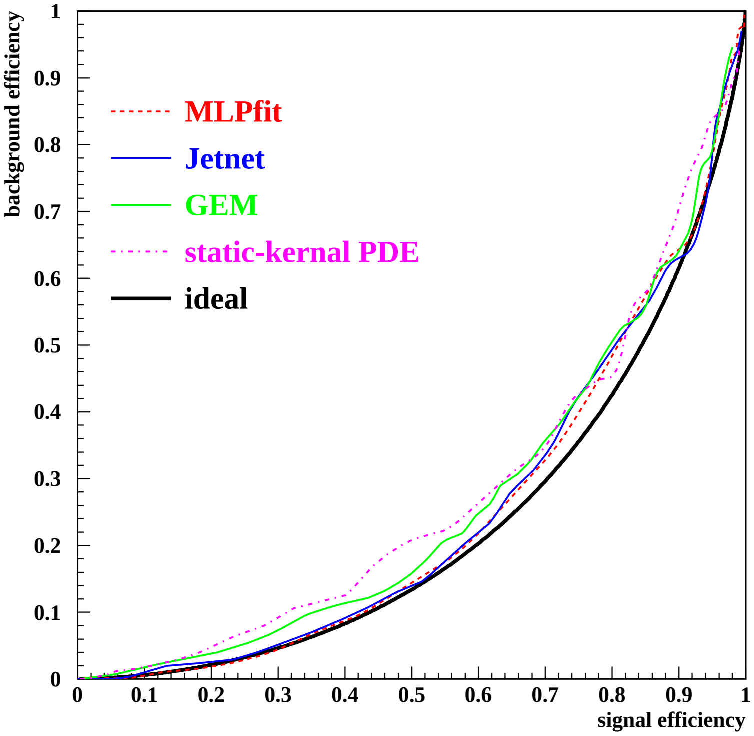

Fig. 2: The same as Figure 1, except that four extra, non-discrimating variables have been added to the analysis. The discrimination power of all the multivariate techniques is significantly degraded by the statistical noise added by these four variables.

1 The remaining variables are added one-at-a-time to the accepted set of variables, and the discrimination statistic is determined for each combination.

2 The variable from step 1 that provides the best improvement in the discrimination statistic is added to the accepted set of variables if it adequately models the behaviour of the data, and if the new discrimination with the added variable is better than the discrimination without that added variable.

Steps 1 and 2 are repeated until no variables pass the acceptance requirement in step 2. The subset of variables found by this method may not be minimal or optimal (but will likely be close). Unfortunately, the number of variables that pass the selection in step 2 is dependent on the size of the training sets used. The next section describes a slightly more complicated algorithm that avoids this problem.

2.2 A Slightly More Complicated Algorithm

This algorithm is very similar to the previous method, except that ’dummy’ variables (which afford no distinction between signal and background) are added to the signal and background Monte Carlo samples. These dummy variables will allow the user to test the “null discrimination hypothesis” (that is to say, they will allow the user to statistically test whether or not an added variable appears to improve the discrimination between signal and background).

The first step in the method is to add N dummy variables to the signal and background Monte Carlo sets. It is important that the same PDF be used to fill the signal and the background dummy variables. The author usually does this by filling the variables with numbers randomly sampled from the Normal distribution. A dozen or so added dummy variables are usually sufficient. These variables are combined with the set of real variables that are believed to perhaps afford some distinction between signal and background.

The user then performs a univariate analysis (using the multivariate analysis method of their choice), determining the $S/\sqrt{S+B}$ discrimination statistic for each of the real variables. The user chooses the one variable that appears to afford the best discrimination between signal and background, and also adequately models the behaviour of the data. This variable forms the nucleus of the accepted set of variables.

Then an iterative process begins:

> **1** The remaining variables are added one-at-a-time to the accepted set of variables, and the discrimination statistic is determined for each combination. The N dummy variables are also added one-at-a-time, and the discrimination statistic is calculated for each. The mean and RMS of the N dummy discrimination statistics will form the basis for the null hypothesis comparison.
>
> **2** The variable from step 1 that provides the best improvement in the discrimination statistic is added to the accepted set of variables if it adequately models the behaviour of the data, and if the new discrimination with the added variable is at least S standard deviations (of the dummy RMS) better than the mean null hypothesis discrimination. S is an arbitrary parameter chosen by the user.

Steps 1 and 2 are repeated until no variables pass the acceptance requirement in step 2. This confidence limit is independent of the sizes of the Monte Carlo sets used to perform the study.

3 Summary

Every variable added to a multivariate analysis adds statistical noise to the measurement. To avoid having this noise wash out the discrimination power of an analysis, it is advantageous to reduce the number of variables to a minimum number of optimal discriminators.

References

[1] DØ Collaboration, V.M. Abazov *et al.*, Phys. Lett. **B517** (2001) 282.

[2] MLPfit, J. Schwindling, `http://schwind.home.cern.ch/schwind/MLPfit.html`

[3] Jetnet 2.0, L. Lonnblad *et al.*, Comput. Phys. Commun. **70** (1992) 167.

[4] B. Knuteson *et al.*, (2001) physics/0108002. Static-kernel PDE methods are also described in “Kernel Probability Density Estimation Methods”, in these proceedings.

[5] The Gaussian Expansion PDE method, developed by the author, is described in “Kernel Probability Density Estimation Methods”, in these proceedings.