PHYSICAL REVIEW D VOLUME 50, NUMBER 9 1 NOVEMBER 1994

Enhancing the top-quark signal at Fermilab Tevatron using neural nets

Ll. Ametller
Departament Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, E-08034 Barcelona, Spain

Ll. Garrido
Departament Estructura i Constituents Matèria, Universitat de Barcelona, E-08028 Barcelona, Spain and Institut de Física d'Altes Energies, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Barcelona), Spain

P. Talavera
Departament Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, E-08034 Barcelona, Spain

(Received 20 May 1994)

We show, in agreement with previous studies, that neural nets can be useful for top-quark analysis at the Fermilab Tevatron. The main features of $t\bar{t}$ and background events in a mixed sample are projected on a single output, which controls the efficiency, purity, and statistical significance of the $t\bar{t}$ signal. We consider a feed-forward multilayer neural net for the CDF reported top-quark mass, using six kinematical variables as inputs. Our main results are based on the exhaustive comparison of the neural net performances with those obtainable from the standard experimental analysis, by imposing different sets of linear cuts over the same variables, showing how the neural net approach improves the standard analysis results.

PACS number(s): 14.65.Ha, 02.50.Sk, 13.85.Qk

The announced evidence of the top quark by the Collider Detector at Fermilab (CDF) [1] at the Tevatron has caused great excitement in the scientific community. Although the statistics are too limited[1] to establish the existence of the top quark, it is however natural to interpret the excess of events as $t\bar{t}$. The experimental situation will certainly improve in the next months and the top quark will hopefully be confirmed. From the theoretical point of view, the consistency of the standard model demands the top to be the partner of the bottom quark, ensuring the absence of flavor-changing neutral currents [3]. The CDF value of the top mass, $m_t = 174 \pm 10^{+13}_{-12}$ GeV [1], is consistent with recent theoretical studies on radiative corrections combined with precision measurements of the Z boson mass and the strong coupling constant at the CERN e^+e^- collider LEP leading to $m_t = 165^{+13+18}_{-14-19}$ GeV [4].

The dominant top production mechanism at the Tevatron is $q\bar{q} \to t\bar{t}$, followed by $gg \to t\bar{t}$. Once produced, the top decays into bW, with the subsequent $W \to l\nu, q\bar{q}'$ decay, in the detector. There are therefore three possible final states for the $t\bar{t}$ signal which, with increasing branching ratios, are (1) two charged leptons, missing energy and two jets, (2) one charged lepton, missing energy and four jets, and (3) six jets.

They need different strategies for top searches and different backgrounds have to be considered respectively. The first channel suffers from a small branching ratio and the presence of two undetected neutrinos that makes top reconstruction unfeasible. It has been analyzed in terms of the correlations among the charged leptons [5] and, recently, it has been suggested to be separable from its possible backgrounds [6]. The most investigated channel so far is the one containing one charged lepton [7]. It has a sizable branching ratio with a moderate background. Still the neutrino escapes detection and hence the event cannot be completely reconstructed. The third channel, six final jets, is the most likely and allows full top reconstruction but at the expense of a huge QCD background. Recently, it has been pointed out that tagging of a b quark can help to obtain acceptable signal-to-background ratios for $m_t < 180$ GeV [8].

All mentioned channels need some specific experimental cuts for detecting jets and/or hard leptons, as well as for their isolation. This, together with detector performance, implies a sensible reduction on the number of possible $t\bar{t}$ candidates, and demands a good efficiency for discerning real from fake $t\bar{t}$ background events. We consider to use neural nets (NN's) for the analysis of experimental data trying to maximize the signal-to-background ratio without significant losses in statistics, in particular to the top analysis at the Tevatron. NN's are by now well known for their ability in pattern recognition giving, after proper training, an approximation to the probability that a given event belongs to some class [9]. They are being used as classifying tools in several high energy applications [10]. Some examples are the Higgs boson search at the CERN Large Hadron Collider (LHC) [11], b and τ analysis [12], quark and gluon jets analysis [13], determination of Z to heavy quarks branching ratios [14], or bottom jet recognition [15]. Some applications of these techniques to the top-quark search have been discussed already in the literature with emphasis on different aspects: In [16] a feed-forward net is trained for $m_t = 100,120,140$ GeV, values that are somewhat below the CDF top mass, relying on PYTHIA [17] for the top quark and the background estimations, thus neglecting spin correlations. The results are favorably compared with the performance of a specific set of linear cuts on several variables, some of which are not used as inputs of the NN, and also compared with the Bayesian decision limit based on probabilities. (In Ref. [18] the results of [16] are

[1]CDF has reported on 12 events, with 6 events for the estimated background, with a 0.26% probability of observing background fluctuation. D0 instead has no clear signal of the top quark [2].

compared with the binary decision tree method.) In Ref. [19] Kohonen's learning vector quantization NN is used and compared to statistical methods and to the Bayesian limit.

In this paper we compare the performance of a multilayer feed-forward NN with the usual procedure for signal-to-background optimization, using exact tree level matrix elements and the CDF reported top-quark mass. (We do not make a comparison with the Bayesian limit since we know from Ref. [9] that NN's are only approximations to such a limit, as pointed out explicitly in [16,19].) In fact, given a set of kinematical variables, we compare the NN result using these variables with the best usual procedure (in fact with many of them) by varying the possible linear cuts over the same variables. We restrict our study at the parton level, without considering detector acceptance, resolution effects, efficiencies, etc., since we are only interested in the performance of the NN versus the usual procedure for enhancing the signal-to-background ratio.

We focused our analysis to the one charged lepton channel

$$p\bar{p}\rightarrow t\bar{t}\rightarrow l\nu jjjj, \tag{1}$$

with $l=e^{\pm},\mu^{\pm}$, using the exact tree level amplitudes with spin correlations [20]. The main background to this process is [21]

$$p\bar{p}\rightarrow Wjjjj\rightarrow l\nu jjjj \tag{2}$$

together with

$$p\bar{p}\rightarrow WW(WZ)jj\rightarrow l\nu jjjj \tag{3}$$

which is an order of magnitude smaller [22]. We have only considered the first mechanism and have used VECBOS[2] [23] for its evaluation.

We have taken $m_t=174$ GeV and have normalized the total $t\bar{t}$ cross section at the Tevatron to 5.1 pb, a value that takes into account $O(\alpha_s^3)$ corrections and the resummation of leading soft gluon corrections to all orders in perturbation theory [24]. CDF measures a $t\bar{t}$ cross section of $13.9^{+6.1}_{-4.8}$ pb [1] which is a factor around 2.5 bigger than the theoretical value we have used. Notice that using the CDF value, the signal-to-background ratio would increase by the same factor. We have used the Harriman-Martin-Roberts-Stirling (HMRS) set 1 structure functions [25] at the scale $Q=m_t$ ($Q=\langle p_t\rangle$) for the top signal (background). We generated events satisfying reasonable acceptance cuts for the jets, charged lepton, and missing transverse momentum,

$$p_T^j, p_T^l, \not{p}_T > 20 \text{ GeV}, \tag{4}$$

and the jets and lepton pseudorapidities

$$|\eta^j|, |\eta^l| < 2, \tag{5}$$

and, requiring jet and lepton isolation,

$$\Delta R_{jl}, \Delta R_{jj} > 0.7, \tag{6}$$

where $\Delta R=\sqrt{(\Delta\eta)^2+(\Delta\phi)^2}$ is the distance in the lego plot. These cuts are intended to simulate the experimental cuts needed to detect jets and hard leptons inside the detector and to select good candidates for top production (from now on these cuts will be referred to as acceptance cuts). The cross section after the acceptance cuts is 0.35 pb (1.2 pb) for $t\bar{t}$ signal (background) in good agreement with Ref. [26]. We generated 4000 $t\bar{t}$ and 4000 background events. The total number of events is essentially limited by the time needed to generate a statistically significant sample for the background. (More efficient generation techniques, which would hopefully circumvent this problem, have been recently proposed [8].)

Notice that the acceptance cuts have to be supplemented either with additional cuts or any other criteria, as a NN for instance, on some kinematical variables in order to assign a single event as signal or background, leading to a reduction of the $t\bar{t}$ and background event samples. (b tagging, for example, reduces the signal by a factor of order 0.3 [27].)

We have considered six kinematical variables in our analysis: (i) $p_T^{W_l}$, the transverse momentum of the leptonically decaying W; (ii) E_T, the total transverse energy; (iii) $m_{W_{jj}}$, the invariant mass of the hadronically decaying W; (iv) m_t, the reconstructed top mass; (v) S, sphericity; (vi) A, aplanarity.

Variables (i) and (ii) are completely defined when assigning the missing transverse momentum to the undetected neutrino. The third variable requires pairing of two jets with invariant mass close to the W mass. Variables (iv), (v), and (vi) need the knowledge of the longitudinal momentum of the neutrino, which is not measured. It can, however, be inferred assuming that the $l\nu$ pair comes from an on-shell W. This leads to a twofold ambiguity which can be resolved to some extent by requiring $t\bar{t}$ reconstruction in the lines suggested by Ref. [26] to which we refer for details. The sphericity and aplanarity, computed for the lepton plus neutrino plus four-jet momenta, take into account the topology of the events expecting larger values from the signal than the background distributions.

The usual strategy for classifying signal- or background-type events is by applying different cuts on the kinematical variables considered, the six mentioned above in our case. These cuts are usually given by simple expressions (for instance, var1 > cut1 and var2 < cut2), so that the different regions are separated by hyperplanes in the variable space (from now on these cuts will be referred to as kinematical cuts). Denoting by T (B) the number of top signal (background) events passing our selection criteria, and T_t the total number of $t\bar{t}$ events selected after the acceptance cuts, Eqs. (4)–(6), one would like to find the best combination of cuts on the kinematical variables such to maximize the efficiency $\eta\equiv T/T_t$ or the purity $P\equiv T/(T+B)$ or both simultaneously. In the latest case, a method could be to maximize the statistical significance of the filtered subsample, $S_s\equiv T/\sqrt{B}$, a criterion that can be used to enhance a new signal from its expected background. In any case, this gives rise to subtle fine-tuning on the cuts to reach the maximization that can become a hard issue for a larger number of kinematical variables considered.

[2]We thank W. Giele for making the VECBOS code available to us.

We are interested in the separation of the signal and background using a layered feed-forward NN which, as we will show, avoids fine-tuning in a multivariable space. A feed-forward NN consists of several layers of units called neurons. Among the layers, we can distinguish one input layer where the information comes in, one or several hidden layers where the information is processed, and one output layer which yields the output of the NN.

The input of the neuron i in layer l is given by

$$I_i^l=\sum_j w_{ij}^l S_j^{l-1}+B_i^l, \quad l=2,3,\ldots, \tag{7}$$

$$I_i^1=in_i^{(e)}, \tag{8}$$

where $in_i^{(e)}$ is the set of kinematical variables for event e, the sum is extended over the neurons of the preceding layer $(l-1)$, S_j^{l-1} is the state of the neuron j, w_{ij}^l is the connection weight between the neuron j and the neuron i, and B_i^l is a bias input to neuron i. The state of a neuron is a function of its input $S_j^l=F(I_j^l)$, where F is the neuron response function. In this Rapid Communication we take $F(I_j^l)=1/[1+\exp(-I_j^l)]$, the so-called "sigmoid function," which is similar to the response curve of the biological neuron and offers more sensitive modeling of real data than a linear function.

The parallel behavior of NN's has the capacity of learning over a set of given examples. A very popular learning algorithm is the error back propagation (BP) [28]. The main objective of the BP is to minimize an error function, also called energy,

$$E=E(\mathrm{in}_i^{(e)},\mathrm{out}^{(e)},w_{kl},B_n)=\frac{1}{2}\sum_e (o^{(e)}-\mathrm{out}^{(e)})^2, \tag{9}$$

by adjusting the w_{kl} and B_n parameters and $o^{(e)}$ being the state of the output neuron, $\mathrm{out}^{(e)}$ its desired state, and e runs over the event sample. Taking the desired output as 1 for each signal event and 0 for each background event, the output of the net, after training, gives the conditional probability that given the observed quantities for a single event, this event is a signal [9], provided that the ratio of signal to background in the learning sample corresponds to the real one.

We have used a three layer NN with six input neurons that are activated with the kinematical variables mentioned in the previous section (normalized to 1 for convenience), a hidden layer with six neurons, and a unique output neuron which desired output is 1 for the signal and 0 for the background. We have found that using six neurons in the hidden layer optimizes the minimum energy.

For the training step we have used 2000 top events and 2000 background events which do not correspond to the expected ratio of cross sections. However, since we are not interested in the conditional probability mentioned above but to study the efficiency and purity as a function of the cut on the output activation of the NN, this fact will not produce any trouble and the learning results will be more efficient. As a test sample, we have taken 570 (2000) top (background) events statistically independent from the training ones. The top/background ratio of the test sample is chosen equal to that obtained from the expected cross sections. All results presented have been obtained from the test sample.

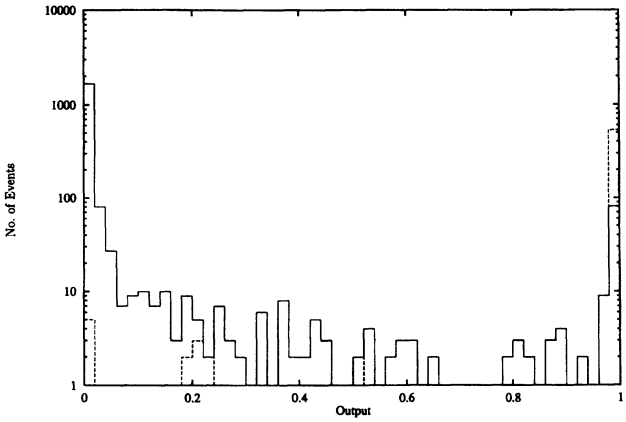

FIG. 1. Distribution of the signal (dashed) and background (solid) events as a function of the NN output activation for the test sample consisting of 570 (2000) top (background) events. Values close to 1 (0) correspond mainly to top (background) events.

Figure 1 shows the distribution of signal and background events as a function of the NN output activation for the test sample. We see two peaks close to 1 and 0 corresponding mainly to the signal and background, respectively. It is clear from this plot that, cutting on the output of the net, we can have samples richer on signal or in background as desired.

The solid (dashed) line in Fig. 2 shows the efficiency (purity) as a function of the net output cut. It is clear that we have to choose an output cut close to 1 if we want high purity or a cut close to 0 for high efficiency. The highest output cut to improve the purity, given a fixed luminosity, would be the one leading to still enough signal events (five at least). This cut will be very close to 1, due to the fact that the efficiency is larger than 0.9 for any value of the output cut except for values very close to 1.

Figure 3 shows the efficiency versus the purity (solid line) when varying the NN output cut from 0 to 0.999 98. The points correspond to some hypercubic cuts applied over the six input variables, and have to be considered as the tradi-

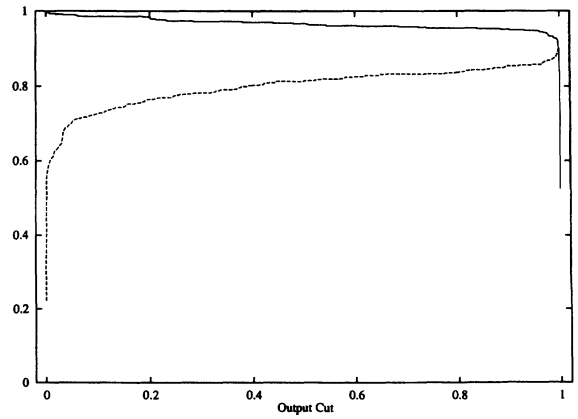

FIG. 2. Efficiency (solid) and purity (dashed) as a function of the NN output cut.

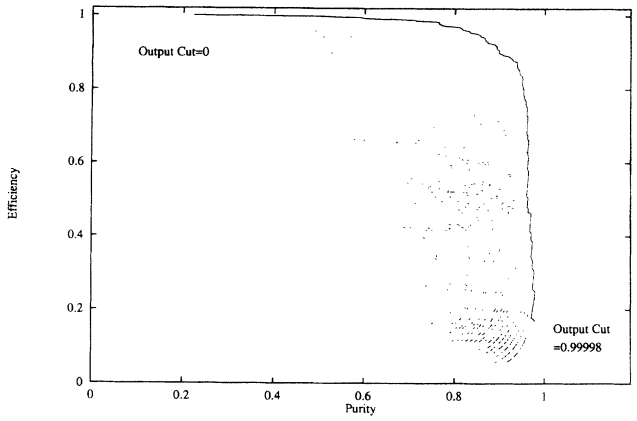

FIG. 3. Efficiency versus purity for the test sample. The solid line shows the NN result whereas the points correspond to several sets of linear cuts (see text) applied to the six input variables.

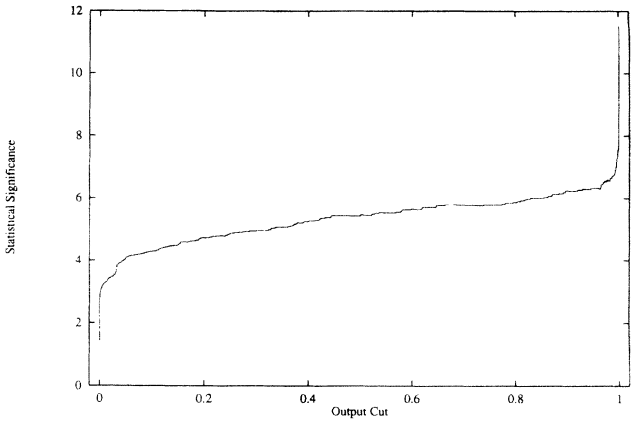

FIG. 5. Statistical significance as a function of the NN output cut for an integrated luminosity of 20 pb^{-1}.

tional procedure (each point represents a given combination of cuts bigger than certain values, or masses located around a certain central value, for instance, $p_t > p_t^{\min}$, $S > S^{\min}$, $m_W - \delta < m_{W_{jj}} < m_W + \delta, \ldots$), chosen favoring the signal in front of the background. We find that the NN performance, working only with one variable, the output of the net, is better than the traditional analysis for any combination of purity and efficiency, showing the great improvement of the method. A complex problem on many variables has been reduced to the study of only one variable, the NN output, which even improves the analysis.

When the important fact is to reveal the existence of the signal, the relevant quantity should be the statistical significance. Values of $S_s > 5$ are commonly accepted as a proof of the existence of a clear signal. Figure 4 shows the relation of the statistical significance versus the efficiency and the purity for $T_t = 1$ signal events (changing the number of signal events, T_t, the surface in Fig. 4 does not modify its shape and only rescales its height which is proportional to $\sqrt{T_t}$). Figure 5 shows the statistical significance as a function of the net output for seven signal events before kinematical cuts (corresponding to an integrated luminosity of 20 pb^{-1}). We see that the statistical significance increases as the output cut increases. As in the case for improving purity, the highest output cut, given a fixed luminosity, would be the one leading to still enough signal events (five at least), and is very close to 1.

One of the problems that is faced in $p\bar{p}$ collisions is the estimation of the background. This seems already to be the case in the present CDF analysis, where the Z+jets and W+jets data are not compatible with theoretical predictions [1]. A factor f of background underestimation could destroy any evidence of the signal. In Fig. 6 we have the allowed region (shaded area) of the output cut versus the factor f of the background ($f = 2$ means that the background is two times bigger as we have computed) where, for the luminosity of 20 pb^{-1}, we still can obtain a 5σ effect with at least five signal events left. Given a fixed factor f the largest and

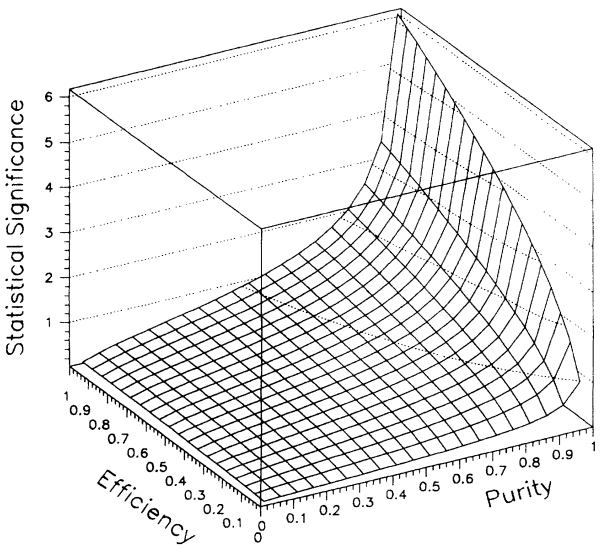

FIG. 4. Statistical significance as a function of the efficiency and purity normalized to $T_t = 1$ signal events. It scales as $\sqrt{T_t}$.

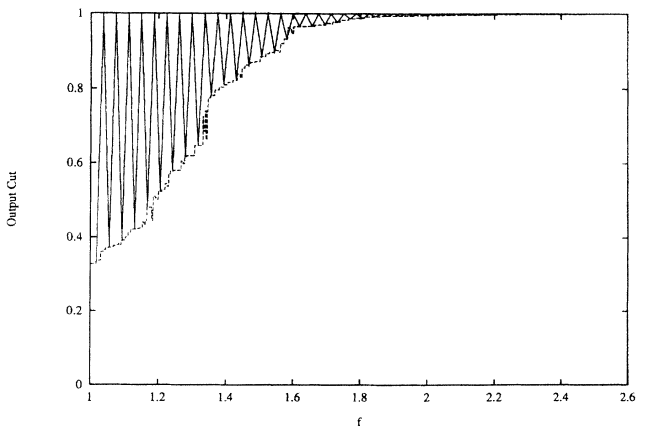

FIG. 6. Allowed region (shaded area) of the output cut versus the factor f of the background ($f = 2$ means that the background is two times bigger as we have estimated) where, for the luminosity of 20 pb^{-1}, we still can obtain a 5σ effect with at least five signal events.

smallest values of the output cut correspond to the highest purity and highest efficiency, respectively. Notice that output cuts very close to 1 are not included in the allowed region, although this is not visible in the plot.

Our results indicate that NN's are suitable for top analysis at Tevatron. Although we focused our study in a particular channel and worked at the parton level, we expect similar behavior for the other channels with the corresponding backgrounds and when performing more realistic analysis including hadronization and detector simulation. We claim neither to have used the best kinematical variables for our analysis, nor to find the best NN topology. Our aim was only to study the potential use of NN's as a cross check to the traditional analysis in terms of cuts on a multidimensional variable space. More elaborated studies are postponed for a forthcoming publication.

In conclusion, we have shown that a NN trained with a mixed sample of $t\bar{t}$ and background events learns the main features of the different samples in a multivariable input space and projects them on a single output. This output turns out to be very useful for discrimination between signal and background events, as it was previously noticed by Baer *et al.* [16]. A comparison among the NN's, trained with a given set of kinematical variables, and a large number of possible combinations of linear cuts, over the same variables, shows that the NN performances (for signal-to-background optimization) are higher than those obtained from the usual experimental procedure for any efficiency and purity. We expect similar behaviors for any set of kinematical variables eventually used. This is indeed partially confirmed by (and can be viewed as a generalization of) the results of Ref. [16], where a NN analysis, using a different set of kinematical variables, was favorably compared with the performance of a unique set of cuts, applied using additional variables in the latter case. Finally, the uncertainty on the background estimation is mapped to a window of possible NN output cuts such that the signal can still be distinguished from the background.

We thank F. del Aguila and G. Stimpfl for discussions. This research was partly supported by EU under Contract No. CHRX-CT92-0004.

[1] CDF Collaboration, F. Abe *et al.*, Phys. Rev. D **50**, 2966 (1994).

[2] D0 Collaboration, S. Abachi *et al.*, Phys. Rev. Lett. **72**, 2138 (1994).

[3] CLEO Collaboration, P. Avery *et al.*, Phys. Rev. Lett. **53**, 1309 (1984); CLEO Collaboration, A. Bean *et al.*, Phys. Rev. D **35**, 3533 (1987).

[4] ALEPH Collaboration, R. Miquel (private communication).

[5] V. Barger, J. Ohnemus, and R. J. N. Philips, Int. J. Mod. Phys. A **4**, 617 (1989).

[6] T. Han and S. Parke, Phys. Rev. Lett. **71**, 1494 (1993).

[7] W. T. Giele and W. J. Stirling, Nucl. Phys. **B343**, 14 (1990); H. Baer, V. Barger, J. Ohnemus, and R. J. N. Philips, Phys. Rev. D **42**, 54 (1990); R. K. Ellis and S. Parke, *ibid.* **46**, 3785 (1992); F. A. Berends, B. Tausk, and W. T. Giele, *ibid.* **47**, 2746 (1993).

[8] J. M. Benlloch, N. Wainer, and W. T. Giele, Phys. Rev. D **48**, 5226 (1993).

[9] Ll. Garrido and V. Gaitan, Int. J. Neural Syst. **2**, 221 (1991); S. Gómez and Ll. Garrido, in *Proceedings of the International Conference on Artificial Neural Networks–94*, edited by M. Marinaro and P. G. Morasso (Springer-Verlag, London, 1994), Vol. 1, p. 549.

[10] V. Gaitan Alcalde, Ph.D. thesis, Universitat Autònoma de Barcelona, 1993.

[11] P. Chiapetta, P. Colangelo, P. de Felice, G. Nardulli, and G. Pasquariello, Phys. Lett. B **322**, 219 (1994); G. Stimpfl-Abele and P. Yepes, Comput. Phys. Commun. **78**, 1 (1993).

[12] M. Joswig *et al.*, DESY Report No. 93-167 (unpublished).

[13] M. A. Graham, L. M. Jones, and S. Herbin, Illinois University Report No. ILL-TH-93-18 (unpublished); A. Vorvolakos, Imperial College Report No. IC-HEP-93-3 (unpublished).

[14] N. de Groot, Ph.D. thesis, Amsterdam University, 1993.

[15] P. Mazzanti and R. Odorico, Z. Phys. C **59**, 273 (1993); K. H. Becks, F. Block, J. Drees, P. Langefeld, and F. Seidel, Nucl. Instrum. Methods Phys. Res. Sect. A **329**, 501 (1993).

[16] H. Baer, D. Dzialo Karatas, and G. F. Giudice, Phys. Rev. D **46**, 4901 (1992).

[17] H. U. Bengtsson and T. Sjöstrand, Comput. Phys. Commun. **46**, 43 (1987).

[18] D. Bowser-Chao and D. L. Dzialo, Phys. Rev. D **47**, 1900 (1993).

[19] A. Cherubini and R. Odorico, Z. Phys. C **53**, 139 (1992).

[20] R. Kleiss and W. J. Stirling, Z. Phys. C **40**, 419 (1988).

[21] H. Baer, V. Barger, and R. J. N. Philips, Phys. Rev. D **39**, 3310 (1989).

[22] V. Barger, T. Han, J. Ohnemus, and D. Zeppenfeld, Phys. Rev. D **41**, 2782 (1990).

[23] F. A. Berends, W. T. Giele, H. Kuijf, and B. Tausk, Nucl. Phys. **B357**, 32 (1991).

[24] E. Laenen, J. Smith, and W. L. Van Neerven, Phys. Lett. B **331**, 254 (1994); R. K. Ellis, *ibid.* **259**, 492 (1991).

[25] P. N. Harriman, A. D. Martin, R. G. Roberts, and W. J. Stirling, Phys. Rev. D **42**, 798 (1990).

[26] V. Barger, J. Ohnemus, and R. J. N. Philips, Phys. Rev. D **48**, 3953 (1993).

[27] S. Protopopescu and G. P. Yeh, presented at the Top Physics Symposium, Madison, Wisconsin, 1992 (unpublished).

[28] P. J. Werbos, Ph.D. thesis, Harvard University, 1974.

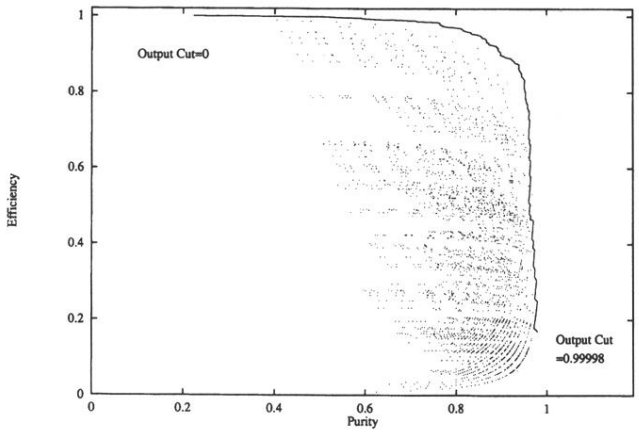

FIG. 3. Efficiency versus purity for the test sample. The solid line shows the NN result whereas the points correspond to several sets of linear cuts (see text) applied to the six input variables.