Rejection of Multi-Jet Background in a Hadron Collider Environment through a SVM Classifier

Federico Sforza, Vittorio Lippi, Giorgio Chiarelli, Sandra Leone

***Abstract*—We optimized a multivariate discriminant software package, based on the Support Vector Machine (SVM) algorithm, to reduce the multi-jet background events in the channel $p\bar{p} \rightarrow e\nu + j\bar{j}$. This channel is important for many physics searches but the multi-jet background can be large and it is difficult to model. We developed a package which allows training set selection, maximization of efficiency and consistency checks. In this paper we will show how the multivariate approach we presented proved to be more efficient compared to the state of art approaches, both in terms of classification accuracy and background contamination**

I. Introduction

In High Energy Physics, since many years, artificial neural networks (ANN) find a number of important applications. Supposed to offer solutions to real-time applications, they have been largely used as offline algorithms. The possibility to appropriately train ANN with simulated sample makes this multivariate approach very appealing. However, especially at hadron colliders like Large Hadron Collider (LHC) of CERN, where there are critical backgrounds to be measured on statistically limited data samples, some limitations of the ANN become evident. In this context we propose to use the Support Vector Machine [1] to classify background based on a *low statistics partly biased* background training sample. This work started in the framework of one of the analyses looking for the Higgs at CDF [2] , and we now present new results based on the improved experience on the use of SVM.

II. Physics Problem

The physics problem we are dealing with is the identification of events containing one electron (e), one neutrino (ν) and 2 jets at the CDF II [3] experiment. The e and the ν are the decay products of a W boson while the jets can derive from interesting processes like Higgs or Single-top decay. The events of interest are produced in the $\sqrt{s} = 1.96$ TeV $p\bar{p}$ collisions at the TeVatron accelerator (Batavia, IL). However, a large background due to multi-jet production, shares the same topology of the signal. Due to the large production probability, its contribution is quite important. The same physics problem can be found in any hadron collider environment, for example at the LHC.

The standard approach for $e\nu$ identification is to use a cut-based definition of the physics objects: the electron is reconstructed as a charged track pointing to a high energy cluster in the electromagnetic section of the calorimeter, with shape compatible with test beam data (electron ID cuts); an imbalance in the total measured transverse energy (Missing E_T or $\not\!\!E_T$), signals the presence of a ν escaping the detector. Although the probability of a multi-jet event faking this topology is low, the large production of multi-jet events makes this is an important background in rare events searches. The nature of fake events is strictly intertwined with the detector operations which is impossible to simulate and it has to be described using data driven models (see next section for details on the selection).

We developed and succesfully used in [2] and in [4] a classifier based on the event kinematic variables and the SVM algorithm to reduce the multi-jet contamination in a sample of $e\nu$ event where the electron is reconstructed in the central region of the CDF detector (CEM). Now we improved our software package and tested the training/rejection algorithm also on electrons reconstructed in the forward region of the detector; here there is a coarser granularity of the calorimeter and only a partial coverage of the tracking chambers, so the probability of multi-jet fakes increase by about a factor of 3 with respect to the CEM selection.

III. Support Vector Machines

The Support Vector Machines (SVM) are supervised training binary classifiers. The SVM algorithm produces the maximum margin hyperplane between the classes of the elements of the training set. Figure 1 shows how the problem can be formalized in the minimization of $|w|^2$ (with w = vector normal to the plane) with the constrain:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \qquad \begin{cases} y_i = +1; & i \in \text{signal} \\ y_i = -1; & i \in \text{bkg} \end{cases} \tag{1}$$

The constrained maximization can be formulated introducing the Lagrange multipliers α as :

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j. \tag{2}$$

This problem has a unique solution with $w = \sum_i \alpha_i y_i x_i$. A penalization on misclassification is added to the objective

Manuscript received November 22, 2011. This work was supported in part by the Universities Research Association (URA), by the Frontier Detectors for Frontier Physics Association and by the Perceptual Robotics Laboratory (PERCRO, Scuola Superiore Sant'Anna, Pisa).

Federico Sforza is with Università di Pisa and INFN (Istituto Nazionale di Fisica Nucleare, Pisa). e-mail: federico.sforza@pi.infn.it

Vittorio Lippi is with Perceptual Robotics Laboratory (PERCRO), Scuola Superiore Sant'Anna, Pisa. e-mail: v.lippi@sssup.it

Giorgio Chiarelli is with INFN (Istituto Nazionale di Fisica Nucleare, Pisa). e-mail: giorgio.chiarelli@pi.infn.it

Sandra Leone is with INFN (Istituto Nazionale di Fisica Nucleare, Pisa). e-mail: sandra.leone@pi.infn.it

978-1-4673-0120-6/11/$26.00 ©2011 IEEE

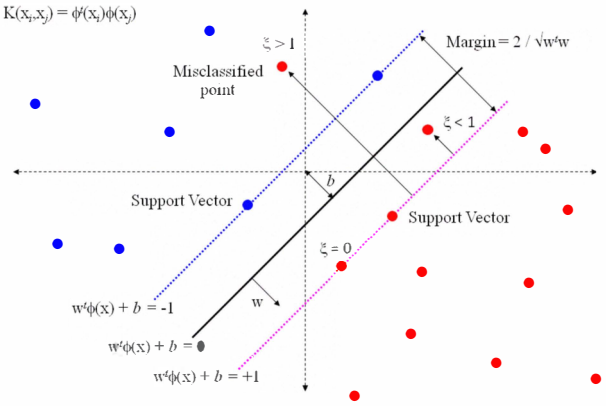

Fig. 1. An example of SVM: two classes of vectors are represented by red and blue dots. The plan leading to a maximum margin separation is defined by the weight vector w and the bias vector b. $\phi : \Re^n \mapsto \mathcal{H}$ maps the points into an higher dimensional space, so to obtain non-linear separation. All the scalar products appear in the form of kernel functions $\mathbf{K}(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$.

function:

$$|w|^2 + C \sum_i \xi_i; \tag{3}$$

subject to:

$$y_i(x_i \cdot w + b) \geq 1 - \xi_i \quad \text{with} \quad \xi \geq 0. \tag{4}$$

The "penalty parameter", C, is an Hyper-parameter to be set before training. Non-linearly separable classes of vectors can be transformed into linearly separable classes by an appropriate function ($\phi(x)$) that maps their elements on a space with higher dimension than the original one. A Kernel function ($\mathbf{K}(x_i, x_j)$) is the composition of the inner product appearing in the Equation 2, with the mapping $\phi(x)$:

$$\mathbf{K}(x_i, x_j) = \phi(x_i) \cdot \phi(x_j), \quad \phi : \Re^n \mapsto \mathcal{H}, \ \mathbf{K} : \Re^n \mapsto \Re \tag{5}$$

$\mathbf{K}$ can be defined without an explicit transformation, just respecting the necessary properties of kernel functions (e.g. see [1])

In this work we use a radial basis function defined by the parameter γ (or Gaussian kernel):

$$\mathbf{K}(x_i, x_j) = e^{-\gamma |x_i - x_j|^2} \tag{6}$$

The corresponding $\phi(x)$ maps to an infinite dimension Hilbert space. The parameter γ is one of the SVM hyper-parameters to be defined before the training.

IV. Training Samples Definition

The SVM algorithm is a supervised classification method: it needs, to perform the training, two sets of events labeled as *signal* or *background*; an original feature of this work is the use of a third validation set composed by *data* with unknown label where a validation cross check of the training is performed. In this paragraph the building procedure of the three sets (signal, background and data) is described.

- **Data:** this is a validation set, it should reproduce the general requirements of the downstream physics analyses where the multi-jet rejection algorithm will be applied. The $e\nu$ forward events online selection [5], from April 18^{th} 2008 to March 6^{th} 2011, requires an electromagnetic cluster with $1.2 > |\eta| > 2.8$ and $E_T > 20$ GeV and an online $\not\!\!E_T > 15$ GeV. The offline selection requires the consistency of the electromagnetic cluster shape with the test beam electron profile, $Iso = \frac{E_T^{0.4}}{E_T} < 0.1$ ($E_T^{0.4}$ sum of the calorimeter energy withing $R = 0.4$ from the electromagnetic cluster), the positive response of the PHoeniX (PHX) tracking algorithm[1], two high energy central jets corrected for detector response ($E_T^{cor} > 20$ GeV, $|\eta| < 2.0$) and $\not\!\!E_T^{cor} > 15$ GeV corrected for the jet reconstruction. The *first* jet, or jet 1, is identified as the one with highest E_T^{cor}, the other jet is called *second*, or jet 2.
- **Signal:** the building of the signal training samples is a fundamental passage to avoid biases in the estimated classification efficiency. By this point of view the forward electron sample requires a special accuracy. As we want to identify $e\nu + 2$jets events, we use a Monte-Carlo (MC) simulation of the $p\bar{p} \rightarrow W \rightarrow e\nu$+jets process calculated with the ALPGEN generator [6] and interfaced with PYTHIA [7] for the final state hadrons simulation. The generator gives us $W + N$ *partons* samples that should be combined to obtain the complete $N-$jets spectrum: we build the final training sample by randomly selecting $W+1, 2$partons events in proportion given by the Leading Order production cross sections ($\sigma_{W+1p} = 225$ pb $\sigma_{W+2p} = 35.3$ pb). Another effect that we need to simulate is the online selection requirements that the CDF experiment uses for the forward electron identification: the trigger is parametrized as a function of electron E_T and the $\not\!\!E_T$ of the event. We used the equation:

$$\varepsilon(x_{e,\nu}) = \frac{1}{1 + e^{-\beta_{e,\nu}(x_{e,\nu} - \alpha_{e,\nu})}}, \tag{7}$$

with:

$$x_e = E_T, \quad x_\nu = \not\!\!E_T; \tag{8}$$

the parameters of the equation, measured on auxiliary samples, are: $\alpha_\nu = 14.57 \pm 0.06$, $\beta_\nu = 0.367 \pm 0.006$, $\alpha_e = 22.34 \pm 0.10$, $\beta_e = 0.41 \pm 0.01$. The training sample has been pruned from the events around the trigger turn-on with the correct proportion. With this procedure we produced a weighted signal sample of about 10^5 events; only $8 \cdot 10^3$ are used in the training while the rest of them are employed in the validation process.
- **Background:** a common method to produce a multi-jet fakes enriched model, is the selection of candidate electrons from data with the same selection but where the electron ID cuts have been reversed. We used $4 \cdot 10^3$ events for the background training set and we kept all the of them ($\simeq 14 \cdot 10^3$) for the validation. The final sample is statistically limited to the data availability and the variables correlated with the inverted cuts may have a bias. These two reasons played a fundamental role in

[1]The complete trajectory of the particle is reconstructed using the inner silicon layers of the tracking volume and the position information of the electromagnetic cluster and of the primary interaction vertex.

our choice of the SVMs as classification algorithm: they are supposed to be optimal on low statistical samples and less prone to over-fitting than, for example, ANN.

Each event of the training samples is defined by a set of 25 input variables. We start with a relatively large set of inputs but we suppose that the final optimal discrimination can be achieved, by the SVM classifier, with a minor number of variables: we leave the duty to prune the set to the optimization algorithm. We choose variables related to the kinematic of the $e\nu + j\bar{j}$ jets events and we tried to avoid the ones directly derived from the electron selection detector information. These are the variables:

- E^e: total reconstructed energy of the electron in the calorimeter.
- E^e_T: projection on the transverse plane of E^e: $E^e_T = E^e \sin\theta$.
- $\not\!\!E^{raw}_T$, or $|\vec{\not\!\!E}^{raw}_T|$: raw missing energy. Imbalance in the sum of the calorimeter energy in the transverse plane.
- *jet 1* E_T,*jet 2* E_T: calorimeter energy of, respectively, the jet with highest E^{cor}_T and the jet with the second highest E^{cor}_T.
- *jet 1 cor*, *jet 2 cor*: correction factor to the calorimeter response for, respectively, the jet with highest E^{cor}_T and the jet with the second highest E^{cor}_T.
- $\not\!\!E^{cor}_T$ or $|\vec{\not\!\!E}^{cor}_T|$: corrected missing energy. $\not\!\!E^{raw}_T$ corrected for the presence of jets and their calorimeter response correction.
- M^W_T: transverse mass of the reconstructed W boson, it is derived as:

$$M^W_T = \sqrt{2(E^e_T \not\!\!E_T - E^e_x \not\!\!E_x - E^e_y \not\!\!E_y)}. \quad (9)$$

- MetSig: missing energy significance, variable related to the goodness of the $\not\!\!E^{raw}_T$ measurement:

$$\text{MetSig} = \frac{\not\!\!E_T}{\sqrt{\Delta E^{jets} + \Delta E^{uncl}}}, \quad (10)$$

where:

$$\Delta E^{jets} = \sum_j^{jets} (cor_j^2 \cos^2\left(\Delta\phi(j, \vec{\not\!\!E}_T)\right) E_{T,j}, \quad (11)$$

$$\Delta E^{uncl} = \cos^2\left(\Delta\phi(E^{uncl}_T \vec{\not\!\!E}_T)\right) E^{uncl}_T, \quad (12)$$

and *uncl* refers to the calorimeter energy not clustered into the electron or the jets.

- $\not\!\!P_T$ or $|\not\!\!P_T|$: missing momentum. Transverse plane imbalance of the sum of the reconstructed charged tracks momenta.
- $\Delta\phi(\vec{\not\!\!P}_T, \vec{P}^e)$, $\Delta\phi(\vec{\not\!\!P}_T, \vec{\not\!\!E}^{raw}_T)$ $\Delta\phi(\vec{\not\!\!P}_T, \vec{\not\!\!E}^{cor}_T)$: angle, on the transverse plane, between the vector of the missing momentum and, respectively, the electron momentum, the raw missing energy and the corrected missing energy.
- $\Delta\phi(\vec{P}^e, \vec{\not\!\!E}^{raw}_T)$, $\Delta\phi(\vec{P}^e, \vec{\not\!\!E}^{cor}_T)$: angle, on the transverse plane, between the electron momentum and, respectively, the raw missing energy and the corrected missing energy.
- $\Delta\phi(\vec{P}^{jet1}, \vec{\not\!\!E}^{cor}_T)$, $\Delta\phi(\vec{P}^{jet2}, \vec{\not\!\!E}^{cor}_T)$: angle, on the transverse plane, between the corrected missing energy and, respectively, the jet 1 and jet 2 momenta.

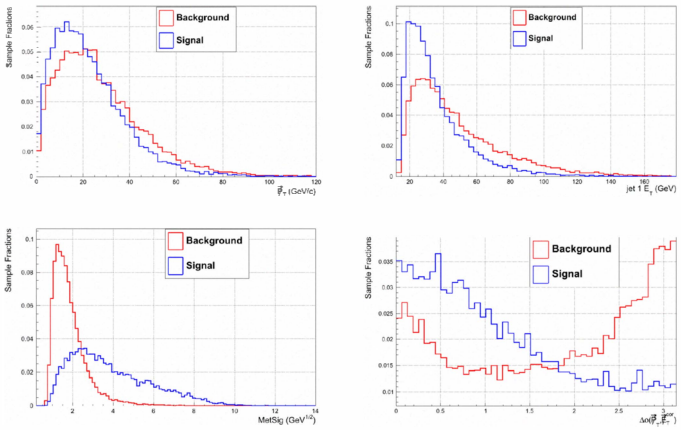

Fig. 2. Example distributions of four input variables for the signal and the background (from left to right, from top to bottom: $\not\!\!E^{raw}_T$, jet 1 E_T, MetSig, $\Delta\phi(\vec{P}^{jet1}, \vec{\not\!\!E}^{cor}_T)$.

- $\Delta R(\vec{P}^{jet1}, \vec{P}^e)$, $\Delta R(\vec{P}^{jet2}, \vec{P}^e)$: angular distance between the electron momentum and, respectively, the jet 1 and jet 2 momenta.
- $\Delta R(\vec{P}^{jet1}, \vec{\not\!\!E}^{cor}_{min})$, $\Delta R(\vec{P}^{jet2}, \vec{\not\!\!E}^{cor}_{min})$, $\Delta R(\vec{P}^e, \vec{\not\!\!E}^{cor}_{min})$: angular distance between the reconstructed momentum of the neutrino[2], where we used the minimal P^ν_z solution, and, respectively, the jet 1, jet 2 and electron momenta.
- $\Delta R(\vec{P}^{jet1}, \vec{\not\!\!E}^{cor}_{max})$, $\Delta R(\vec{P}^{jet2}, \vec{\not\!\!E}^{cor}_{max})$: angular distance between the momentum of the neutrino, reconstructed as in the preceding point, where we used the maximal P_z solution, and, respectively, the jet 1 and jet 2 momenta.

Figure 2 show some example distribution of input variables for background and signal.

V. Variable Selection and Consistency Checks

Variable selection proved to improve the performance of SVM classifiers [8], especially when the variables have a different relevance in the discrimination algorithm [9]. Furthermore, a reduced set of variables makes further tests shorter and, from a broader point of view, identifying an optimal subset of variables gives information about their importance in the analyzed process.

We implemented a framework to aid the selection of a subset of the variables. To achieve all this we need to evaluate the performance of the classifier in each given configuration, following the flowchart of Fig. 3

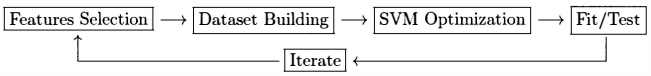

Fig. 3. Flowchart of feature selection - training - test procedure.

[2]The P^ν_z of the neutrino can not be measured, so we infer it from the W boson mass using the electron momentum and setting $\vec{P}^\nu_T = \vec{\not\!\!E}_T$. P^ν_Z is the solution of a quadratic equation, which may have two real solutions, one real solution, or two complex solutions: for the case with two real solutions we distinguish the minimal and the maximal one. For the complex case, the real part is chosen.

TABLE I
DEFINITION OF CONFUSION MATRIX.

Signal classified as *Signal*	*Background* classified as *Signal*
Signal classified as *Background*	*Background* classified as *Background*

The performances of each configuration of variables are tested using the best SVM defined by the parameters C and γ that give the highest classification accuracy on the training set. In this context the performance is defined by two figures of merit: the accuracy achieved by the classifier in the cross validation and the background contamination returned by a bi-component fit, over signal MC and background model, on the $\not\!\!E_T$ distribution in data.

The *confusion matrix* of the discriminant (Table I) is a standard tool in machine learning classification studies: it shows the percentage of events correctly assigned or not on samples of known label. In our case, for each configuration we run the trained SVM on the full background sample and on the full signal Monte-Carlo sample.

The second point is one of the key features of this work: the background model reproduces most of the distributions of the real background but the distributions fail when they are correlated with the inverted quality cuts used to create the QCD enriched sample. It is fundamental to have a cross check to identify mis-modeled variables. To do that we exploit a standard methodology for the evaluation of multi-jet background normalization: we fit the $\not\!\!E_T$ distribution of data with templates of signal and background (they account for $\gtrsim 90\%$ of the data sample). Figure 7 shows the result of the fit on the complete samples before any multi-jet rejection. We apply the same procedure also after the application of the SVM on *data*, *signal* and *full background* samples. The fit allows us to define two quality criteria for the considered SVM: we mark the variable configuration as BAD if:

- the χ^2 of the fit is greater than 5 (the requirement is loose because we expect other processes might slightly influence the shape);
- **or** the fraction of mis-identified background is not consistent between the results of the fit and the confusion matrix.

Notice that the quality of the fit is not directly optimized by the SVM training, so we are performing a consistency check of our classifier in an unbiased sample (data) with an independent technique.

A. Grid Evaluation of the SVM

In order to perform the described operations we developed the software to train the SVM and produce the needed consistency checks. Our software analyzes a given number of combination of variables taken from the 25 of our starting sets. Results are then displayed in an interactive scatter plot (Fig. 5) where we show background contamination derived from the fit on y-axis and signal efficiency from MC on x-axis. Then, it is possible to select any desired configuration and, keeping it fixed, add other n-combinations of the remaining variables.

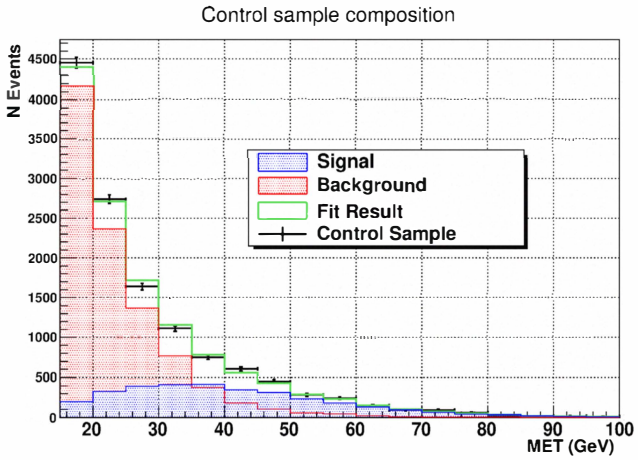

Fig. 4. Bi-component fit on the $\not\!\!E_T$ variable on the data sample with background and signal templates. The fit represents the initial composition of the forward $e\nu+2$jets sample with a multi-jet contamination of $f_{Bkg} \simeq 64\%$ (in the region of interest $\not\!\!E_T > 25$ GeV).

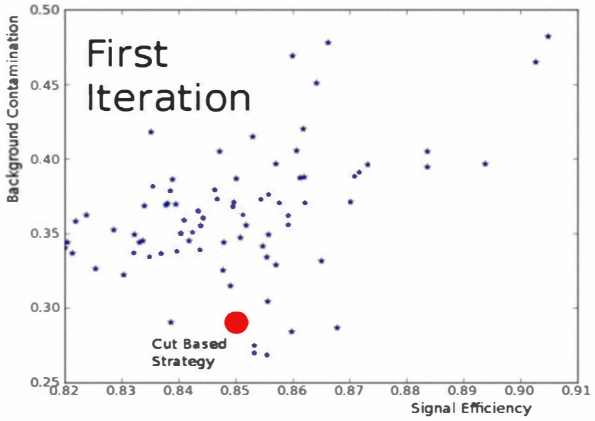

Fig. 5. Different SVM configuration are displayed on a "signal-efficiency vs background-contamination" scatter plot. This plot shows the SVM performances resulting from the combination of 3 out of 25 input variables; the red dot shows the performance of the cut-based strategy.

An extensive research over all the possible combinations of variables is unfeasible due to the huge number of combinations: in our case with a total of 25 variables, we have 33554431 combinations. To allow a feasible scan of most of the combination phase space we started with all the configuration given by 3 variables, then we added, to the *best result*, the combinations of 3 extra variables, then we added 4 variables, again, all the possible combinations, to the two best combinations. After this step the *best result* did not improve any more. The background contamination *vs* signal efficiency scatter plots after the three iterations are displayed in Figure 5 and Figure 6. All this required to explore just $C(25,3)+C(22,3)+2C(19,4) = 5778$ combinations and we can assume that it represent a good, if not optimal, encoding of the independent information among the complete set of input variables.

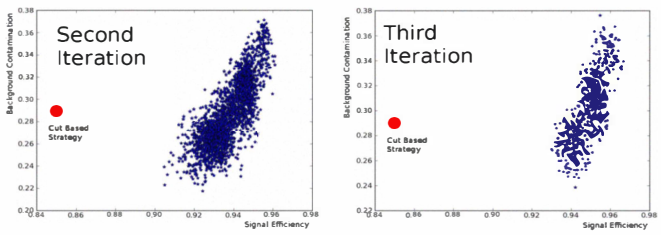

Fig. 6. *Right:* SVM performances resulting adding combination of 3 variables to the best configuration obtained from the first iteration of the SVM training (Figure 5); the red circle shows the performance of the state of art cut based selector. *left:* SVM performances resulting from the configuration selected by adding 4 other variables to the best configuration obtained in the right plot, the efficiency and background rejection power do not improve any more in a relevant way.

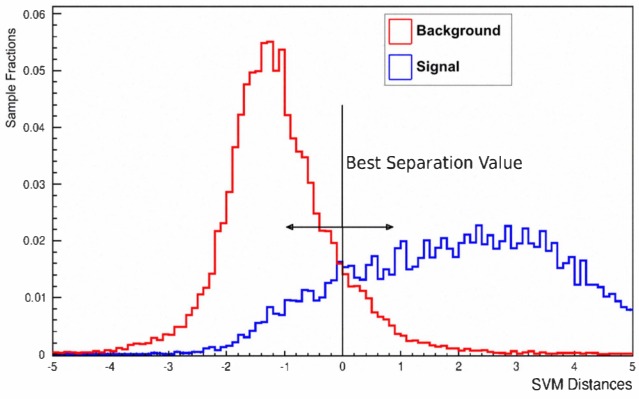

Fig. 8. Distributions of SVM *output* function for the signal and the background samples.

VI. Results and Conclusions

The best achieved result is a 10-variables SVM configuration exploiting E_T^e, E^e, $\not\!\!E_T^{raw}$, $\not\!\!P_T$, jet 1 *cor*, jet 2 *cor*, MetSig, $\Delta\phi(\vec{P}^e, \not\!\!\vec{E}_T^{raw})$, $\Delta R(\vec{P}^{jet1}, \not\!\!\vec{E}_{min}^{cor})$, $\Delta R(\vec{P}^{jet2}, \not\!\!\vec{E}_{min}^{cor})$. Figure VI shows the result of the $\not\!\!E_T$ fit in this configuration: the allowed multi-jet contamination is $f_{Bkg} \simeq 24\%$ and the signal efficiency is $\varepsilon_{Sgn} \simeq 95\%$. This result can be compared with the initial background contamination $f_{bkg} \simeq 64\%$ and with the best available (in the CDF collaboration) multi-jet rejection method that gives $f_{Bkg} \simeq 29\%$ and $\varepsilon_{Sgn} \simeq 84\%$.

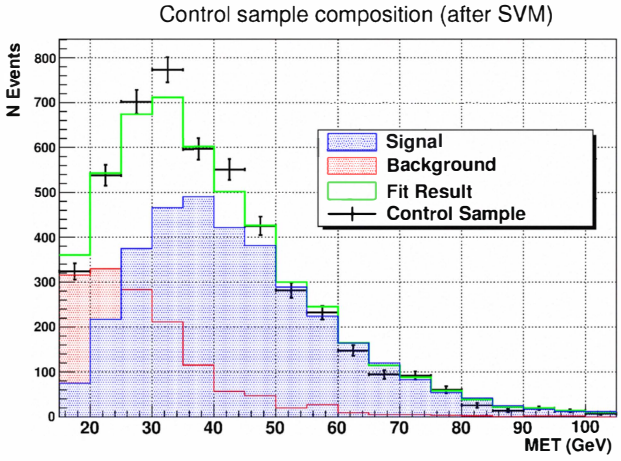

Fig. 7. Bi-component fit on the $\not\!\!E_T$ variable on the data sample with background and signal templates after the application of the best SVM model available: the multi-jet contamination is reduced to $f_{Bkg} \simeq 24\%$ (in the region of interest $\not\!\!E_T > 25$ GeV) with a signal efficiency of $\varepsilon_{Sgn} \simeq 95\%$.

Finally, another relevant advantage of out software is that we can access directly the SVM output function:

$$\sum_i \alpha_i y_i K(x_i, x) + b \tag{13}$$

Figure 8 shows this variable before the classification is performed. Although the SVM is a binary classifier, this new variable can be used to move the classification threshold into sub-optimal values maybe useful in analysis where a higher or lower multi-jet contamination is acceptable.

We can conclude that the multivariate approach we presented proved to be more efficient compared to the state of art approaches, both in terms of classification accuracy and background contamination

Acknowledgment

We thank the Perceptual Robotics Laboratory (PERCRO, Scuola Superiore Sant'Anna Pisa) and Prof. Carlo Avizzano, from Scuola Superiore Sant'Anna, for the support given to Vittorio Lippi.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, October 2007.

[2] G. Chiarelli, V. Lippi, and F. Sforza, "Rejection of multi-jet background in $p\bar{p} \to e\nu + jj$ channel through a svm classifier," in *Journal of Physics: Conference Series, CHEP*, 2010.

[3] D. Acosta, "it measurement of the tt production cross section using lepton+jets events with secondary vertex b-tagging," *Phys. Rev. D*, vol. 71, no. 052003, 2005.

[4] F. Sforza, "Diboson search and multivariate tools in the $l\nu + b/\ c$ jets channel at cdf," in *Nuovo Cimento C - Colloquia on Physics, IFAE*, 2011.

[5] The Cartesian right-handed coordinate system has the z-axis oriented along the proton beam direction and the y-axis poiting North, the xy-plane is called *transverse* and the quantity prjected on it are noted with a subscript T. A polar coordinate system $(r,\ \phi,\ \theta)$ can be defined with origin in the center of the detector: the angle θ measured from the z-axis and the angle ϕmeasured in the transverse plane starting from the x-axis. We define: the pseudorapidity $\eta = -\ln\tan(\theta/2)$ and the angular distance $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

[6] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, "ALPGEN, a generator for hard multiparton processes in hadronic collisions," *JHEP*, vol. 07, p. 001, 2003.

[7] T. Sjstrand, L. Lnnblad, L. Tp, and S. Mrenna, "Pythia 6.2 - physics and manual," 2001.

[8] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for svms," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 668–674.

[9] Y. wei Chen, "Combining svms with various feature selection strategies," in *Taiwan University*. Springer-Verlag, 2005.