

OPEN ACCESS

Rejection of Multi-jet Background in $p\bar{p} \rightarrow e\nu + j\bar{j}$ Channel through a SVM Classifier

To cite this article: Federico Sforza *et al* 2011 *J. Phys.: Conf. Ser.* **331** 032045

View the [article online](#) for updates and enhancements.


You may also like

- [WMAP-compliant benchmark surfaces for MSSM Higgs bosons](#)
John Ellis, Thomas Hahn, Sven Heinemeyer et al.
- [Mixed higgsino dark matter from a large SU\(2\) gaugino mass](#)
Howard Baer, Azar Mustafayev, Heaya Summy et al.
- [Supersymmetry discovery potential of the LHC at \$s^{1/2} = 10\$ and 14 TeV without and with missing \$E_T\$](#)
Howard Baer, Vernon Barger, Andre Lessa et al.




The
Electrochemical
Society

Advancing solid state &
electrochemical science & technology



DISCOVER
how sustainability
intersects with
electrochemistry & solid
state science research



Rejection of Multi-jet Background in $p\bar{p} \rightarrow e\nu + j\bar{j}$ Channel through a SVM Classifier

Federico Sforza¹, Vittorio Lippi², Giorgio Chiarelli¹

¹ INFN Pisa, Largo B. Pontecorvo 3, 56127 Pisa, IT.

² PERCRO Lab - Scuola Superiore Sant'Anna, Via Martiri 11, 56127 Pisa, IT

E-mail: sforza@fnal.gov, v.lippi@sssup.it, giorgio.chiarelli@pi.infn.it

Abstract. We test and optimize a multivariate discriminant software package, based on the Support Vector Machine (SVM) algorithm, to reduce the multi-jet background events in the channel $p\bar{p} \rightarrow e\nu + j\bar{j}$. We use the CDFII data-set, collected at the TeVatron $p\bar{p}$ collider, where this channel provides the signature for many important physics processes: e.g. associated Higgs production, WZ, single top events. The *Multi-jet* background can be large and difficult to reject but, in this paper, we show that an appropriately trained SVM can handle it in an effective way. The developed programs perform training set selection, efficiency maximization and consistency checks; we also discuss the robustness of the discriminant. A classification accuracy $\geq 95\%$ can be reached using Monte Carlo simulated signal and a data-driven background model (limited by statistic) with a background rejection of $\simeq 90\%$.

1. Introduction

Our signal consists in a W boson, decaying into $e\nu$, and two jets. We use the Support Vector Machine (SVM) algorithm to reduce the multi-jet (also named “QCD background”) contamination, where the electron and the neutrino are faked by mis-measured jets.

We developed a software package, based on LibSVM [1] library, able to perform algorithm training, variable ranking, signal discrimination and robustness test.

Despite the small probability of a jet faking e and ν , the large cross section of multi-jets events at hadron colliders makes this process a dangerous background to many physics searches. As we are interested in rare processes (eg. W/ZH , single top) the selection must retain a large efficiency for the signal: we set our threshold to 95%.

2. Physics Problem

We apply the SVM discriminant to a sub-sample of the CDFII [2] data-set, where we identify the physics channel $p\bar{p} \rightarrow e\nu + j\bar{j}$ requiring the following selection [3]:

- one Tight Isolated Central Electron (CEM): i.e. a good quality, high P_T track matched to a significant energy deposit ($E_T > 20$ GeV) in the central electromagnetic calorimeter, with low activity in the nearby area of the calorimeter ($Iso < 0.1$);
- exactly two jets reconstructed with $|\eta| < 2.0$ and $E_T > 20$ GeV;
- missing transverse energy ($\cancel{E}_T > 20$ GeV) as signature of the escaping neutrino.

Multi-jet events can pass the same requirements, if one of the jets fakes the electron and the \cancel{E}_T is either mismeasured by the detector, faked by misidentified (or undetected) minimum ionizing particles or produced by neutrinos associated with decay of heavy quarks.

The selection is applied to both signal and background training samples (see below) and to the data we want to classify. Data corresponds to a $\int \mathcal{L} dt \simeq 4.3 \text{ fb}^{-1}$ collected by a high- P_T electron trigger.

2.1. Training Samples

We built our training-set using 8000 signal events and 4000 background events (to emulate the data composition):

signal: $W + 2\text{partons}$ Alpgen Monte-Carlo [4], where the W is forced to decay into electron and neutrino. We have $\simeq 10^5$ generated events and we keep $\simeq 9 \times 10^4$ events as a control sample (i.e. not used for training).

background: due to the nature of the background (a mixture of physics processes and detector response), there is no simulated models that can be trusted to provide the accurate description needed for training. Therefore we use a data-driven approach to obtain a suitable sample: we select events with a fake electron by reversing some of the “electron quality” requirements (at least 2 out of the 5 cuts), used to identify the shape of the electromagnetic shower in the calorimeter. This selection is named “anti-electron” and produces a QCD enriched sample which is, however, limited to a few thousands of events. The sample can be flawed by mis-modeling in some of the variable correlated to the shape variables of the electron.

3. Support Vector Machines

The Support Vector Machines (SVM) are supervised training binary classifiers. In our problem we can rely only on a *low statistics, partly biased background model*: SVM are designed to offer a possible solution to these issues.

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:

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

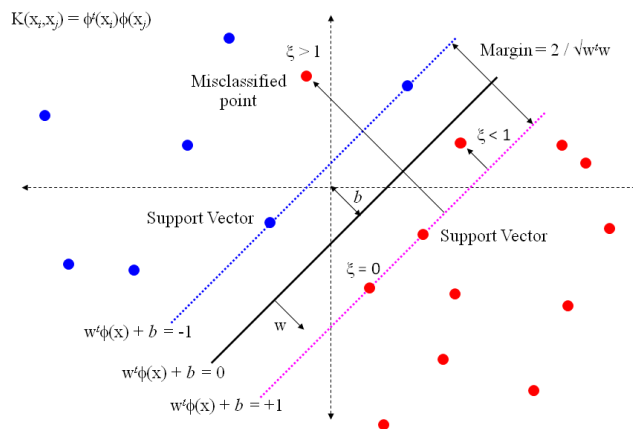


Figure 1. An example of SVM: two classes of vectors are represented by red and blue dots. The plan leading to a maximum margin separation is defined by the weight vector w and the bias vector b . $\phi : \mathbb{R}^n \mapsto \mathcal{H}$ maps the points into an higher dimensional space, so to obtain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

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

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

This problem has a unique solution with $w = \sum_i \alpha_i y_i x_i$.

To take in account the possibility that the training set is not separable a penalization on misclassification is added to the modulus of the w vector:

$$|w|^2 + C \sum_i \xi_i; \quad (3)$$

subject to:

$$y_i(x_i \cdot w + b) \geq 1 - \xi_i; \quad (4)$$

$$\xi \geq 0. \quad (5)$$

The “penalty parameter”, C , is an Hyperparameter to be set before training.

3.1. Kernel Methods

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 ($\mathbf{K}(x_i, x_j)$) automatize this, being the composition of the inner product appearing in the Equation 2, with the mapping $\phi(x)$:

$$\mathbf{K}(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad \text{with } \phi : \mathbb{R}^n \mapsto \mathcal{H}, \quad \mathbf{K} : \mathbb{R}^n \mapsto \mathbb{R} \quad (6)$$

\mathbf{K} can be defined without an explicit transformation, just respecting the necessary properties of kernel functions (e.g. see [5])

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

$$\mathbf{K}(x_i, x_j) = e^{-\gamma |x_i - x_j|^2} \quad (7)$$

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

4. Variable Selection and Robustness Algorithms

We extended the features input system implemented in LibSVM to perform the best variable selection: a reduced set of features improves the robustness of the classifier, makes further tests leaner (see paragraph 4.2) and, from a broader point of view, identifying an optimal subset of variables gives information about their importance in the analyzed process.

To achieve all this we need to evaluate the performance of the classifier in each given configuration, following the flowchart of Figure 2

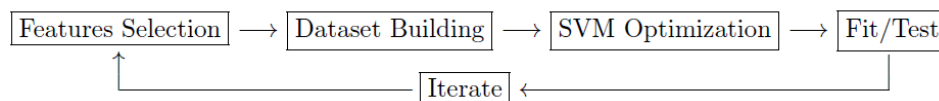


Figure 2. Flowchart of feature selection - training - test procedure.

The performances of each configuration of variables is tested using the best SVM defined by the parameters C and γ that give the higher classification accuracy on the training set. In

Table 1. Definition of confusion matrix.

<i>Signal</i> classified as <i>Signal</i>	<i>Background</i> classified as <i>Signal</i>
<i>Signal</i> classified as <i>Background</i>	<i>Background</i> classified as <i>Background</i>

this context the performance is defined by two figures of merit: the confusion matrix of the discriminant and the background contamination returned by a bi-component fit (signal MC + background model) on the \cancel{E}_T distribution in data.

The *confusion matrix* of the discriminant (Table 1) 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. After the application of the SVM on *data*, *signal* and *full background* samples, we fit the \cancel{E}_T distribution in data with templates of signal and background (they account for $\gtrsim 90\%$ of the data sample). 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 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. An example of the fit is shown in Figure 5.

The procedure we described in this section has been implemented for this specific work. Although we have not yet developed an automatic configuration system, all the software is easily adaptable to other background rejection problems.

4.1. Grid Evaluation of the SVM

As already stated, we aim to the best classification accuracy using the smallest possible set of variables. In this way we can easily evaluate systematics and keep track of the physics meaning of the features. In order to fulfill this program we developed the software needed to perform training and consistency checks for a given number of combination of variables taken from the 22 of our starting sets. Results are then displayed in an interactive scatter plot (Figure 3) 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.

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 22 variables and 4 different composition of the training sets, we have 16777212 combinations. We tried all the configuration given by 1, 2 and 3 variables, then we added, to the *best results*, the combinations of 2 and 3 extra variables. After this step the *best result* did not improve any more.

The *best result* is a 6-variables configuration that exploits the maximal uncorrelated information among the input variables. In Figure 3 is marked by a red circle.

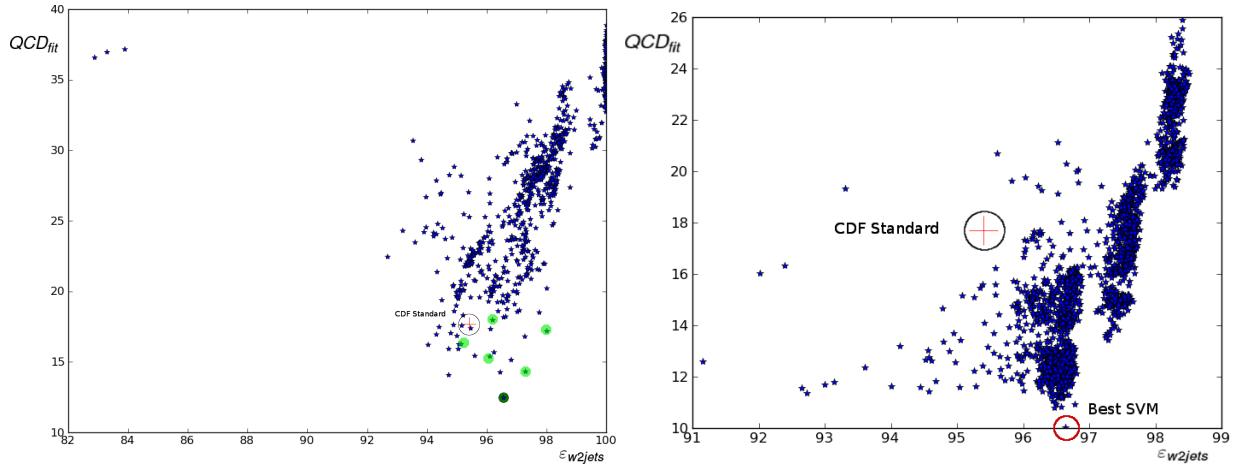


Figure 3. Different SVM configuration are displayed on a “signal-efficiency vs background-contamination” scatter plot. *Left:* SVM performances resulting from the combination of 1, 2 and 3 out of 22 input variables; the circled cross shows the performance of the cut-based strategy; the green dots are the configurations selected to add more variables. *Right:* SVM performances resulting from the configuration selected by the green dots in the left plot and adding 1, 2, 3 other variables; the red circled dot shows the best configuration found.

4.2. Discriminant Robustness

We tested the robustness of the classifier by checking how the SVM accuracy is affected by the uncertainty on the input variables. We performed a test modifying each variable by a $\pm 10\%$ and keeping the others unchanged. This is a standard procedure for neural network based classifiers in the context of particle physics and it can be considered a benchmark. Figure 4 shows a pictorial example of that with two variables. In Table 2 we show the robustness test for the 6 variable configuration considered as the best one.

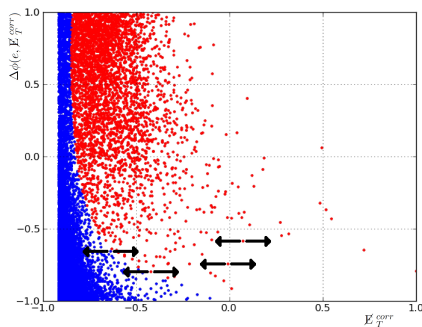


Figure 4. Visual example of the classification (*blue*: background, *red*: signal) produced by a SVM configuration featuring two input variables ($\Delta\phi(e, \vec{E}_T^{corr})$ and \vec{E}_T^{corr} , both scaled in $[-1, 1]$). Classification stability can be estimated by the number of correctly classified vectors after a variation of the input variables.

5. Results

Our algorithm produced an optimal SVM using 6 variables as input features: Lepton P_T , \vec{E}_T^{cor} , $\Delta\phi(e, \vec{E}_T^{raw})$, Jet2 E_T^{raw} , Jet2 E_T^{cor} and the MetSignificance (a variable that relates \vec{E}_T with jet corrections). We use the training set with a $\vec{E}_t > 15$ GeV to enhance the signal-like background component. The best SVM configuration is shown as red circled dot in the signal efficiency *vs* background contamination diagram (Figure 3) with $f_{Bkg}^{Data} \simeq 10\%$ given by the fit on data,

Table 2. Change in the classification efficiency varying the input variables by 10%.

var	$\varepsilon -$	original	$\varepsilon +$
Lep P_t	91.4%	94.5%	92.6%
\cancel{E}_t^{cor}	86.4%	94.5%	87.4%
Jet2 E_t^{cor}	92.4%	94.5%	94.1%
MetSig	88.0%	94.5%	90.0%
$\Delta\phi(e, \cancel{E}_t^{raw})$	94.2%	94.5%	94.3%
Jet2 E_t^{raw}	92.4%	94.5%	94.1%

$\varepsilon_{Sgn}^{MC} \simeq 97\%$, from MC. We can compare this result with the state of art cut-based strategy: $f_{Bkg}^{Data} \simeq 18\%$ and $\varepsilon_{Sgn}^{MC} \simeq 95\%$.

Besides the physics result itself our main achievement was the development of a flexible software able to perform training, variable selection, validation, and to chose the preferred SVM depending on maximum purity or maximum signal efficiency. Figure 5 shows how a high purity SVM configuration works on the background reduction.

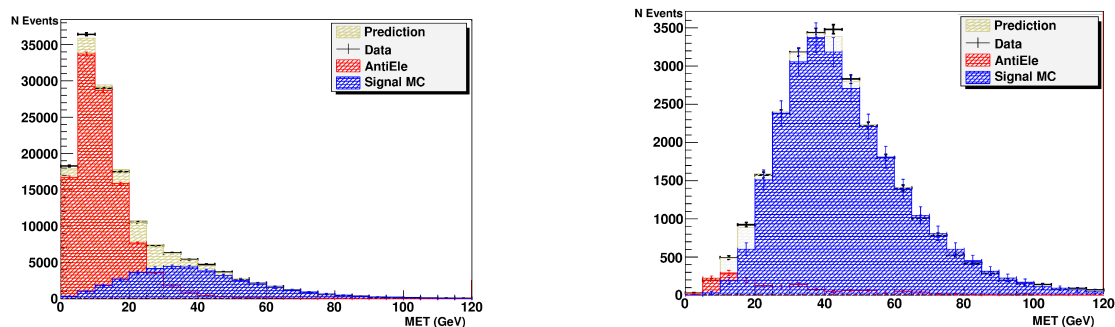


Figure 5. Result of the bi-component fit on Missing Transverse Energy (MET) shape after the application of a SVM specifically selected to have high background rejection: data sample (crosses) as the composition of multi-jet background (red) and $W + 2$ jets signal samples (blue) before (Left:) and after (Right:) the application of the SVM discriminant.

Acknowledgments

We thank the Universities Research Association (URA) for the support given to Federico Sforza.

References

- [1] Chang C C and Lin C J 2001 *LIBSVM: a library for support vector machines*
- [2] D Acosta *et al* 2005 *Phys. Rev. D* **71**
- [3] The cylindrical coordinate system has its origin in the center of the detector with θ, ϕ the polar and azimuthal angles. We define: $\eta = -\ln \tan(\theta/2)$, $P_T = P \sin \theta$, $E_T = E \sin \theta$, $\cancel{E}_T = |-\sum_i E^i \hat{n}^i|$ (\hat{n}^i unit vector pointing to the i^{th} calorimeter tower), $Iso = \frac{\sum_i E_T^i}{E_T}$ with E_T^i summed in a cone of $R = \sqrt{\Delta\eta^2 + \Delta\phi^2} = 0.4$ around E_T . The superscript cor indicates the quantities corrected for the energy response of the detector.
- [4] Mangano M L, Moretti M, Piccinini F, Pittau R and Polosa A D 2003 *JHEP* **07** 001
- [5] Bishop C M 2007 *Pattern Recognition and Machine Learning (Information Science and Statistics)* 1st ed (Springer) ISBN 0387310738