Published by Institute of Physics Publishing for SISSA

Received: *May 16, 2006*
Accepted: *June 21, 2006*
Published: *July 24, 2006*

Applying rule ensembles to the search for super-symmetry at the large hadron collider

Jan Conrad

The Royal Institute of Technology (KTH), Physics Department
AlbaNova University Center
10691 Stockholm, Sweden, and
PH-Department, CERN
CH-1211, Geneva 23, Switzerland
E-mail: conrad@particle.kth.se

Fredrik Tegenfeldt

Iowa State University
Ames, IA 50011-3160, U.S.A.
E-mail: Fredrik.Tegenfeldt@cern.ch

Abstract: In this note we give an example application of a recently presented predictive learning method called *Rule Ensembles.* The application we present is the search for supersymmetric particles at the Large Hadron Collider. In particular, we consider the problem of separating the background coming from top quark production from the signal of supersymmetric particles.

The method is based on an expansion of base learners, each learner being a rule, i.e. a combination of cuts in the variable space describing signal and background. These rules are generated from an ensemble of decision trees. One of the results of the method is a set of rules (cuts) ordered according to their importance, which gives useful tools for diagnosis of the model. We also compare the method to a number of other multivariate methods, in particular Artificial Neural Networks, the likelihood method and the recently presented boosted decision tree method. We find better performance of *Rule Ensembles* in all cases. For example for a given significance the amount of data needed to claim SUSY discovery could be reduced by 15 % using *Rule Ensembles* as compared to using a likelihood method

Keywords: Supersymmetry Phenomenology, Statistical Methods.

Contents

1. Introduction

Multivariate methods are widely used in high energy physics, for a review, see for example [1] and references therein. A typical task is the classification of a data set into two classes, often signal and background. With few exceptions, the concept of a multivariate method can be formulated as follows: The goal is to find the function F for which

$$y = F(\vec{x}) \tag{1.1}$$

where y is a target value and $\vec{x}$ denotes the vector of input values. In particle physics applications with the task of separating signal from background, y encodes signal and background class (for example $+1$ for signal and -1 for background), and $\vec{x}$ is a set of observables (for example the transverse momentum of particles or reconstructed mass) used to distinguish between these two classes. Finding this function is a minimization problem where

$$F_{\mathrm{opt}}(\vec{x}) = \arg\min_{F} E_{\vec{x}y} L(y, F(\vec{x})) \tag{1.2}$$

Here $E_{\vec{x}y}$ denotes the expectation value over the distribution of all variables in the data to be characterized. $L(y, F(x))$ is the cost function, i.e. the penalty for predicting a value y when a different value y_{true} is the true value. Different multivariate methods differ in the way the cost function L is defined and how the function F is constructed. Commonly the approximation to F is found by applying some learning algorithm to a set of known cases, a "training" sample. In this note we will describe and apply a learning method which is based on the idea of "ensemble learners", called *Rule Ensembles (RE)*, which has been presented recently by Friedman [2, 3].

The paper is organized as follows: in the next section we will give a brief introduction to the method of ensemble learning machines and the particular implementation of *RE* considered here. For more details on this method we refer the reader to [2] and [3]. In section 3 we will describe the search for super-symmetric particles and in particular rejection of top background. In section 4 we show how the method performs on the problem considered, followed by a comparison with other common methods in section 5. The final section is devoted to conclusions.

2. Ensemble learning machines

Ensemble learning machines are constructed by building linear combinations of learning machines (in that case called "base learners"). Assuming that the prediction value of the m^{th} base learner is given by $f_m(\vec{x})$ then the prediction value of the ensemble is given by

$$F(\vec{x}) = a_\circ + \sum_{m=1}^{M} a_m f_m(\vec{x}) \tag{2.1}$$

$f_m(\vec{x})$ could be for example the prediction of the m^{th} decision tree in an ensemble of M decision trees. The minimization problem thus becomes the problem of finding the coefficients in this linear expansion in base learners.

A recently published example an ensemble learning machine (though the authors did not present in the context of ensemble learners) is the method of boosted decision trees [4] which in that example were shown to perform better than Artificial Neural Nets (ANN).

2.1 Rule based ensembles

The base learners presented in this note are of the form:

$$r_m(\vec{x}) = \prod_{j=1}^{n} I(x_j \in s_{jm}) \tag{2.2}$$

where I(·) indicates the truth of its argument and s_{jm} indicates a subset of all possible values of variable x_j. Thus, $r_m(x)$ is either 1 or 0 depending on whether or not x_j falls into its s_{jm}. We will call $r_m(x)$ a *rule* in the remainder of this paper. A typical rule in context of particles physics could for example be ($n = 1$):

$$r_m(N_{el}) = I(2 < N_{el} < 5) \tag{2.3}$$

i.e., the number of electrons is larger than 2 and smaller than 5. The variables considered could be both numerical (usually the case in physics applications, like the above example) or categorical (for example, job status ∈ {tenured, post-doc,unemployed}). With rules as base learners, equation (2.1) writes

$$F(\vec{x}) = a_\circ + \sum_{m=1}^{M} a_m r_m(\vec{x}) \tag{2.4}$$

For completeness, we mention that linear functions of the input variables are particularly difficult to approximate by such a set of rules, which is why the particular implementation of *RE* presented here includes optionally a term which explicitly includes a dependence on the input variables $\vec{x}$:

$$F(\vec{x}) = a_\circ + \sum_{m=1}^{M} a_m r_m(\vec{x}) + \sum_{j=1}^{n} b_j l_j(x_j) \tag{2.5}$$

where $l_j(x_j) = \min\left(\delta_j^+, \max(\delta_j^-, x_j)\right)$. The $\delta_j^{+,-}$ functions are the β and (1-β) quantiles of the data distribution (taken over the N events for each x_j), respectively. These functions assure that in case of outliers instead of the outlier a quantile of the distribution is chosen. Values for β are generally small, $\beta \sim 0.025$.

2.2 Ensemble generation and Minimization

The particular algorithm to generate the ensemble and to find the coefficients can differ. In the case presented in this paper, the algorithm used is based on the importance sampled learning ensemble (ISLE) methodology described in [6]. Here we restrict ourselves to give a short summary only. For a given set of base learners $\{r_m(\vec{x})\}_1^M$ a linear regularized regression is performed on the training data.

$$\{\hat{a}\}_0^M = \arg \min_{\{a_m\}_0^M} \sum_{i=1}^{N} L\left(y_i, a_0 + \sum_{m=1}^{M} a_m r_m(\vec{x}_i)\right) + \lambda \sum_{m=1}^{M} |a_m| \tag{2.6}$$

where we omit the linear term for simplicity. The first term in equation 2.6 is the previously mentioned measure of the prediction risk (see equation 1.2) and the second term is the “regularization” term which penalizes large values of the coefficients for each base learner. The ensemble of rules is generated using a decision tree. Each decision tree is viewed as a collection of rules and used as a base learner $r_m(\vec{x})$. In order to understand a decision tree it is best to describe how it can be constructed. The first step is to select one of the input variables and loop over all events. For each event value of the discriminant variable divide the training sample into two, one containing all events with a value less than the considered value and one with all events having a value larger than the considered value. The event value which gives best separation between signal and background is chosen and the procedure is repeated for all discriminant variables considered. The discriminant variable and its event value giving overall best separation is used to form the first decision in the decision tree. The training sample is thus divided into two new samples (*branches*) corresponding to the ones produced by this decision. This procedure is then repeated for each branch in turn, recursively until a predetermined number of final branches is obtained or until the branch consists of pure signal or background. For a more detailed description of decision trees we refer the reader to for example [7].

From equation (2.6) it can be seen that there is the need for defining a cost function. The cost function chosen in the example presented in this note is:

$$L(y,F) = [y - \min(-1, max(1,F))]^2 \tag{2.7}$$

which has been shown to perform comparable to other commonly used methods (e.g. [8]).

Calculations presented in this note have been performed within the *R* statistics framework [9] using the RuleFit package presented in [2]. For the comparison between multivariate methods, the software package TMVA [10] was used.

3. Search for super-symmetric particles in ATLAS

3.1 Introduction

Super-symmetry (SUSY) [11] is an extension of the Standard Model. It is particularly attractive since it provides a solution to the the so-called hierarchy problem and also offers the possibility to incorporate gravity into a quantum field theory. SUSY postulates the existence of superpartners to all known particles. The superpartners of bosons are fermionic (spin $n/2$, where n is an integer) whereas the fermion partners are bosonic (spin n). SUSY particles have a quantum number called R-parity. If this quantum number is conserved, there is consequently a stable SUSY particle which will not decay further, the lightest stable particle (LSP). Experimental studies of SUSY are complicated by the fact that a large amount of free parameters are added to the SM, the minimal extension having 106 free parameters. SUSY is therefore often (and also in this note) studied in a limit which is simplified at the Planck scale (so-called SUGRA) [18, 19]. In this approximation, the number of extra free parameters to be considered reduces to 5. The remaining parameters are the universal scalar masses m_0, the gaugino masses $m_{1/2}$, a trilinear coupling A_0, the ratio $\tan\beta$ and a sign sgnμ.

The mass scale of this theory is dependent on the electroweak breaking scale which is given by the Higgs mass. Currently, the experimental lower limit of the Higgs mass is 114.4 GeV/c^2 [19] and from theoretical arguments, its upper limit is about 800 GeV/c^2. Consequently, SUSY could be discovered at the TeV scale, thus being quickly accessible at the Large Hadron Collider (LHC) [21]. The LHC is a proton-proton collider currently being

Model	σ_{LO} (pb)	m_0 (GeV/c^2)	$m_{1/2}$ (GeV/c^2)	A_0	$\tan\beta$	$sgn\mu$
SU1	6.8	70	350	0	10	+
SU3	19.3	100	300	-300	6	+

Table 1: SUGRA parameters and leading order cross-sections for models used in this paper.

Source	σ (pb)	Characteristics
QCD	$1.5 \cdot 10^9$	Small $\not{E}_T$ but very large cross-section
Top	580	Multijet and large $\not{E}_T$ (escaping hard neutrinos) events
W+jets	1200	Similar to top but $\not{E}_T$ less hard

Table 2: Main background sources to SUSY. Although top has the smallest cross section, it is the hardness of $\not{E}_T$ (and therefore more SUSY like) which makes top likely to be the largest background.

built at the European Center for Nuclear Research (CERN) outside Geneva. The envisaged center of mass energy in proton-proton collisions is 14 TeV and the design luminosity will be $\mathcal{L} = 10^{-2}pb^{-1}s^{-1}$. At the LHC, there will be 4 multi-purpose experiments: ATLAS [12], CMS [13], LHCb [14] and ALICE [15]. ATLAS and CMS focus on the discovery and precision measurements of the Higgs sector and SUSY. LHCb is designed for B-physics. Part of the LHC program includes collisions of heavy ions (lead), which are the main objective of ALICE. In this paper, we will focus on SUSY studies using the ATLAS detector.

3.2 The samples

The ATLAS collaboration considers several points in SUGRA parameter space. In this paper we will only consider two of them, referred to as SU1 and SU3. Their SUGRA parameters are summarized in table 1.

The detector response for each sample has been obtained through a full detector simulation using GEANT4 [16].

3.2.1 Signal

Irrespective of SUSY model, signatures are complex multijet events possibly involving hard leptons and heavy quark jets. In the models considered here, R-parity is conserved and therefore each SUSY event produces stable heavy and weakly interacting particles that escape undetected. One of the main characteristics for these types of event is therefore large missing transverse energy, $\not{E}_T$.

3.2.2 Background

Several background sources are currently considered by the ATLAS collaboration. An overview of the main sources is given in table 2.

The relevant decay chain in top events is $t \rightarrow bW$ where the W can either decay hadronically or leptonically. Leptonically decaying W is particularly difficult to distinguish from SUSY since it is characterized by large $\not{E}_T$ from escaping neutrinos. This is the topology considered in this paper.

Discriminator	Description
$\not{\!\!E}_T$	Missing E_T
M_{eff}	$\not{\!\!E}_T + \sum_{\mathrm{jet}} p_T^{\mathrm{jet}}$
p_T^0, p_T^1	First two hardest jets
H_T	$\sum_{\mathrm{jet}} p_T^{\mathrm{jet}}$
H_T^{3j}	$H_T - p_T^0 - p_T^1$
H_E	$\sum_{\mathrm{jet}} E_{\mathrm{jet}}$
(A)planarity, sphericity	Combinations of eigenvalues of jet momentum tensor
N_{jets}	Number of jets
N_{bjets}	Number of b-jets
N_e, N_μ	Number of electrons/muons

Table 3: The discriminating variables used in the separation of SUSY signal from top background

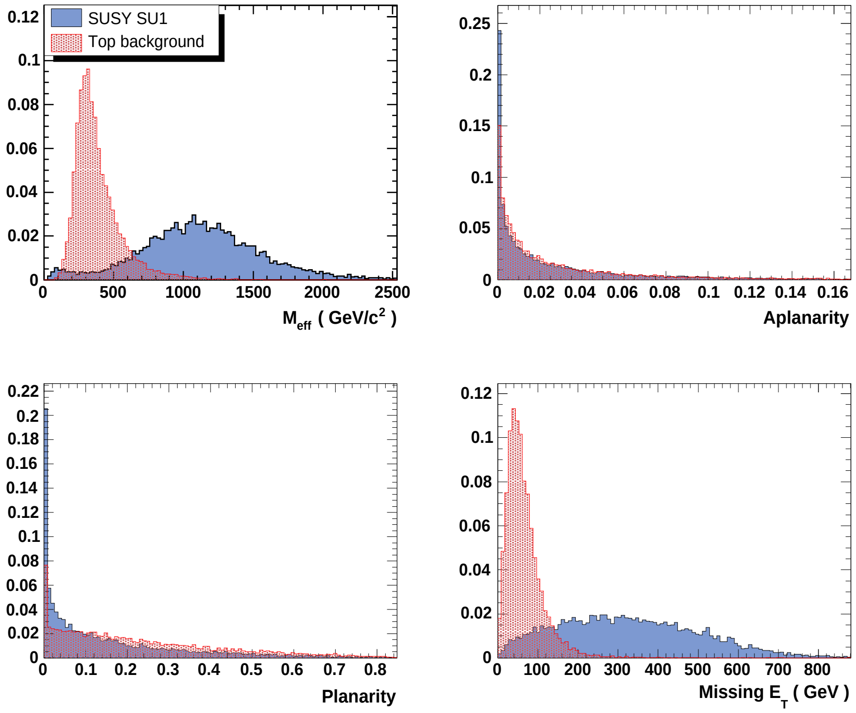

Figure 1: Some example plots of discriminators used in the separation of SUSY signal from top background

3.2.3 Discriminators

In this study we considered fourteen discrimination variables. Except for the previously mentioned $\not{\!\!E}_T$, we considered various combinations of jet momenta as well as the b-jet content. Top events will predominately have 2 b-jets whereas in SUSY several topologies for 0,1,2 or more b-jets are possible. A summary of the discriminators used is given in table 3. Some example plots are presented in figure 1.

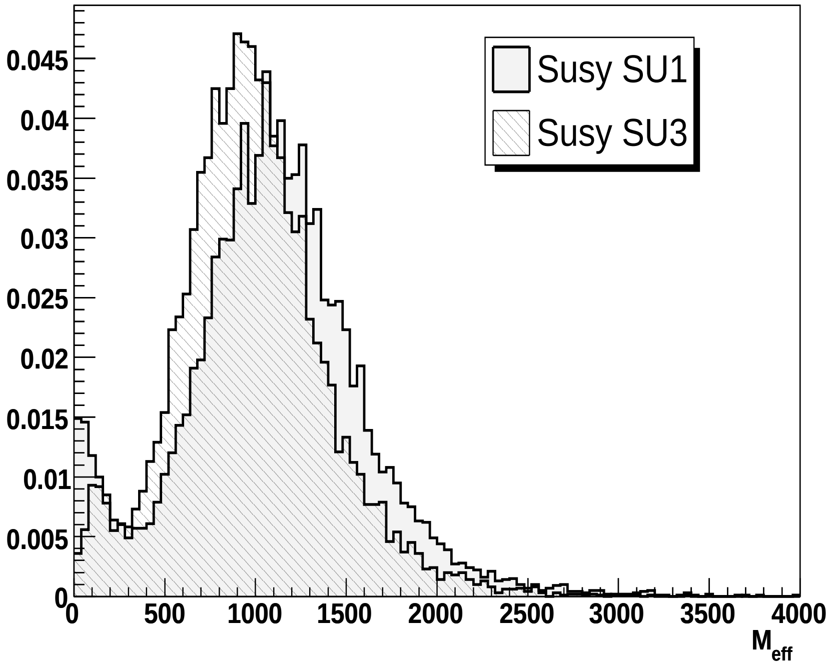

Figure 2: The discriminator M_{eff} (defined in table 3) for the SU1 and SU3 benchmark models

The difficulties constraining SUSY models results constitutes an unavoidable systematic effect: the SUSY model used to train the learning machine might not be the one nature has chosen. Training and testing on two different SUSY models thus yields information about what the effect of training on the wrong model might be as well as give us information about how stable RE are under systematic shifts of the input variables. The difference in one of the discrimination variables, M_{eff} (defined in table 3) can be seen in figure 2. The distribution is shifted by about 20 %.

4. Performance of the method

The performance of the method can be measured in different ways. We will here present two ways of assessing the performance. Firstly, cross validation is used to estimate the absolute error rate, average error rate and area under the ROC curve. The ROC curve is just the correlation between negative error rate (i.e. class signal is misclassified as background) and positive error rate (i.e. class background is misclassified as signal). Cross validation is a common method in which the training sample is randomly subdivided into sub-samples on which the model is mutually trained and tested. Table 4 compares the results for the model which was trained on SU1 SUSY model only and one which was trained on a mix between SU1 and SU3. Also included are the results for a model where we only use the four most important variables of the original sample (see section 4.1.2).

Model	1-A_{ROC}	neg. error rate	pos. error rate	aver. error rate
SU1 (14 variables)	0.024	0.064	0.082	0.073
MIX (14 variables)	0.029	0.072	0.092	0.082
SU1 (4 variables)	0.026	0.068	0.085	0.076

Table 4: Error rates and area under the ROC curve for the samples considered in this study

From this table, it can be seen that the rate of top events misclassified as SUSY events is slightly higher than vice versa. It can also be seen that the method only becomes slightly worse. For example the average error rate increases by 0.3 % in case only the four most important variables are used. Worst performance of the method (though also only slighlty worse) is obtained if the signal sample is in itself consisting of two distinct classes (in this case SU1 and SU3).

Of more practical interest for particle physicists is the performance of the method on independent test samples. In particular score distributions give an intuitive impression on how good the methods perform. We show the obtained scores for the model trained on SU1 and tested on independent test samples consisting of SU1 events, SU3 events and top events in figure 3.

These score distributions can be used to calculate efficiency for signal and background as a function of cut value in the score. In figure 4 we show the correlation between background and signal efficiency for the combinations (trained on SU1, tested on SU1 and SU3 as well as trained on the mixed sample and tested on SU1 and SU3).

Best performance is non-unexpectedly found in the case that we train on SU1 and test on SU1 (SU1-SU1). Worst performance is found if we train on a mixture of SU1 and SU3 and test on SU3. The effect of a systematic shift in signal modeling can be seen from the difference of the curve where we trained on SU1 but tested on SU3, which gives the next to worst result. For example at a signal efficiency of 80 % the background contamination increases by roughly 50 % (from $\sim$ 2 % to $\sim$ 3 %), still a relatively moderate increase. An interesting case is the case where we train on a mixture of SU1 and SU3 but test on SU1. Here, for large signals efficiency behaves equally well as the case (SU1-SU1), whereas it approaches the case where we train on SU1 and test on SU3 for small signal efficiencies.

4.1 Diagnostics

In this section we illustrate how diagnostics of this method can be used to arrive at a simpler, still almost as efficient characterization of the data.

4.1.1 Rule importance

The importance of the rules is defined by the absolute value of their respective coefficients, $|\hat{a}_k|$ and their support, s_k. For the rules this is expressed as

$$I_k = |\hat{a}_k \cdot \sqrt{s_k(1 - s_k)}| \,. \tag{4.1}$$

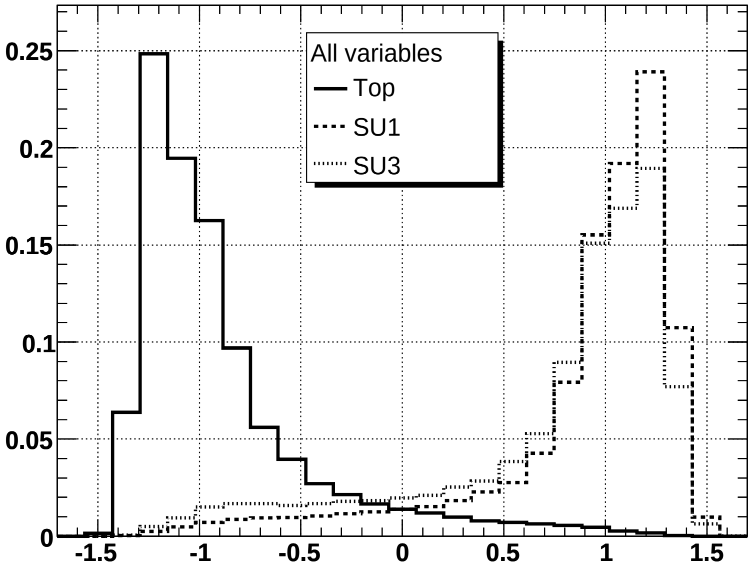

Figure 3: Scores for the case training on SU1 sample and testing on SU1, SU3 and top sample.

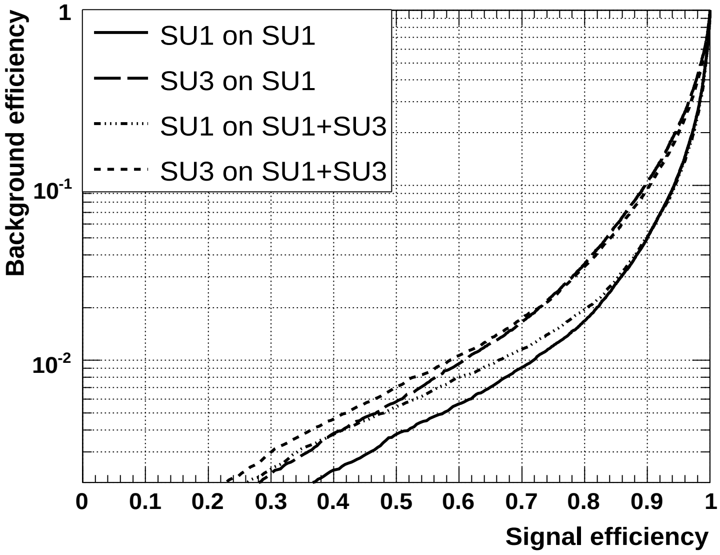

Figure 4: Background vs. signal efficiency for the different samples, trained on SU1 and an equal mixture of SU1+SU3 and tested on SU1 and SU3, respectively.

The support for each rule is defined as its average response of for the given data.

$$s_k = \frac{1}{N}\sum_{i=1}^{N} r_k(\vec{x}_i)\,. \tag{4.2}$$

Rule		Coefficient	Relative Importance
1	$246.8 < \not\!\!E_T < 2180$	0.1539	100.0
2	$20.5 < p_{T,2}^{\text{jet}} < 236.2$ $\not\!\!E_T < 145.5$	-0.1371	96.1
3	$142.5 < M_{\mathit{eff}} < 738.6$ $\not\!\!E_T < 205.3$	-0.1369	95.8
4	$139.4 < M_{\mathit{eff}} < 681.5$ $\not\!\!E_T < 193.5$	-0.1215	85.2

Table 5: The four most important rules using all variables trained on SU1 signal.

For the linear predictors the importance becomes

$$I_j = |b_j| \cdot std(l_j(x_j)) \tag{4.3}$$

where $std()$ is the standard deviation over data. We summarize the four most important rules in table 5.

The cut in $\not\!\!E_T$ is (not unexpected) the most important cut, followed by a combination of $\not\!\!E_T$ and the p_T of the next-to hardest jet. While in the case of SUSY search this is a almost trivial statement, in more complicated cases this might be valuable information by giving hints on which variables should be paid special attention to and in which region of the parameter space. For example, one could find out that the cuts are performed in a region of the parameter space where the quality of the variable (say for example, resolution in p_T) is expected to be worse than in other regions. An improvement of the quality of the variable in that region could then be attempted.

4.1.2 Input variable importance

Another important diagnostic of the model is the relative importance of the input variables to the model. This figure is connected to the rule importance since variables which participate in the most important rules will be judged more important than those participating in less important rules. The importance will also be determined by the frequency in which the variables participate in the rules. The corresponding measure is written as

$$J_l = I_l + \sum_{x_l \in r_k} I_k / m_k \tag{4.4}$$

where m_k is the number of variables participating in rule r_k. The ten most important input variables are shown in figure 5. From this figure it is clear that the most important variables are $\not\!\!E_T$, M_{eff}, and p_T of the hardest and the next to hardest jets.

One possibility then arises by constructing a much simpler model only including these most important variables. Indeed, the area of the ROC curve becomes only negligibly smaller (see table 4). The corresponding efficiency curves calculated from testing on independent test samples are compared in figure 6.

From figure 6 it can be concluded that the model including only the four most important variables is performing only marginally worse than the model including all fourteen

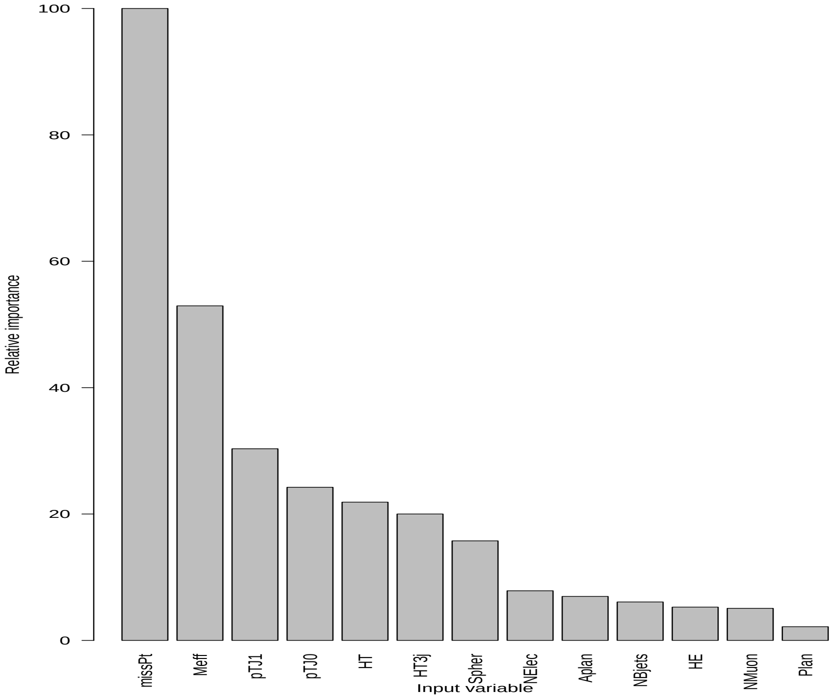

Figure 5: The importance of the variables participating in the sample

discriminating variables. In particular, the degradation in performance is negligible compared to the degradation which is introduced by training on the “wrong” super-symmetric model.

5. Comparison with other methods

In general, it is difficult to compare different multi-variate methods, since it is not simple to prove that in all cases optimal parameters where used. We therefore assume the “out of the box” approach, i.e. we compare different methods using recommended default values. For *RE* all results presented so far are essentially with default parameters of the machine, i.e. with a minimal amount of tuning. In this section we use the package TMVA [10] to compare *RE* with other common methods. In particular, we consider the Likelihood method, an Artificial Neural Network, the H-Matrix method, Boosted Decision Trees and the Fisher Discriminant. The following subsections are devoted to a brief description of each method. The vector $\vec{x} = \{x_0, x_1, \ldots, x_{N-1}\}$ denotes the N different discriminating variables used. $F(\vec{x})$ denotes, as before, the test statistic used for the separation between the classes. The final subsection will describe the results of the comparison.

5.1 Description of the different methods

5.1.1 Likelihood

For each event we can define a likelihood for it to be a signal and a background event

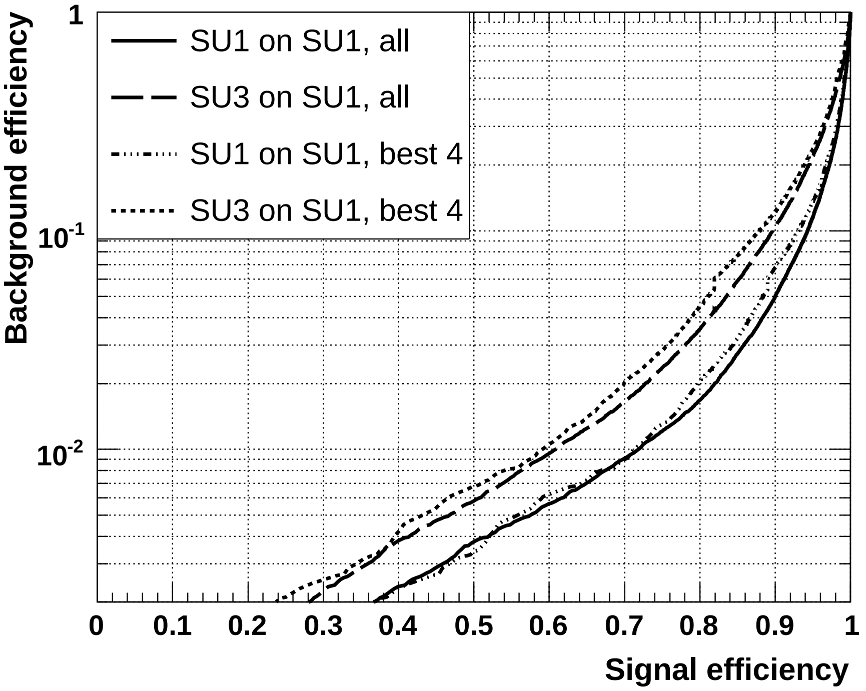

Figure 6: Background vs. signal efficiency for the model using only the four most important variables as compared to the model using all variables

through the product of the probability densities of each discriminating variable

$$\mathcal{L}(\vec{x}) = \prod p_i(x_i) \tag{5.1}$$

The probability density functions are determined using a training sample. As a test statistic, the likelihood ratio is used

$$F(\vec{x}) = \frac{\mathcal{L}_S(\vec{x})}{\mathcal{L}_S(\vec{x}) + \mathcal{L}_B(\vec{x})} \tag{5.2}$$

where the subscripts S and B denotes signal and background class, respectively. For this method it is assumed that all discriminators are uncorrelated (i.e. independent given the class label). Linear correlations can be included by first diagonalizing the correlation matrix. The likelihoods can then be constructed in the transformed variable space where the correlations are minimal.

In the comparison presented below, the method assuming uncorrelated discriminators is used.

5.1.2 H-matrix

In this method, χ^2 estimators are obtained for the signal and background samples

$$\chi^2(\vec{x}) = (\vec{x} - \vec{\bar{x}})^T C^{-1} (\vec{x} - \vec{\bar{x}}) \tag{5.3}$$

The test statistics is then constructed from the ratio

$$F(\vec{x}) = \frac{\chi^2_S(\vec{x}) - \chi^2_B(\vec{x})}{\chi^2_S(\vec{x}) + \chi^2_B(\vec{x})} \tag{5.4}$$

5.1.3 Fisher Discriminant

The general form of a Fisher test statistic is a linear combination of the discriminating variables

$$F(\vec{x}) = \sum_i a_i x_i \tag{5.5}$$

where the coefficients $\{a_i\}$ are selected such that $F(\vec{x})$ yields maximum separation between the signal and background classes. In order to determine a_i, the average of each variable and the variable covariance matrix for the signal class and ditto for background are used. In this paper we use the so called Mahalanobis-Fisher discriminant where the covariance matrix between the classes is included. The coefficient a_i is a measure of the discriminating power of the corresponding variable.

5.1.4 Artificial Neural Network

The Fisher discriminant works well for separated classes. If the discriminators are not linear, a more general parameterizations of the test statistic can be considered. In particular, for the case of a so called multi-layer perceptron, the test statistic is then formed by a linear combination of *nodes* organized in *layers*. In the input layer, the nodes are given by the input variables. The output layer consists of one or more nodes corresponding to the various output classes. In between these layers, an arbitrary number of *hidden* layers with an arbitrary number of nodes can be used. Each node in a given layer is a function of all nodes in the previous layer. The test statistic takes the form

$$F(\vec{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i h_i(\vec{x})\right) \tag{5.6}$$

where $s(\cdot)$ is the *activation function*, often set to $(1+e^{-x})^{-1}$. The functions $h_i(\vec{x})$ are of the same form as $F(\vec{x})$ but act on the previous layer. Since a neural network can contain an arbitrary number of layers and nodes, it is possible to make more or less good designs. The chosen design in this paper is a network with 4 input variables and 2 hidden layers with 4 nodes each.

5.1.5 Boosted decision trees

Boosted decision trees are in detail described elsewhere [4], here we will constrain ourselves to a short description. Boosted decision trees is another example of an ensemble learner. The base learner is in this case a decision tree and boosting is the algorithm used to construct the ensemble. We gave a description of decision trees in section 2.2. The main idea of the *boosting* algorithm consists in a re-weighting of misclassified events. Consider a decision tree constructed in the way described above. Training events which are classified as the opposite class will get an increased (*boosted*) weight and then a new tree is built

Method	sig. eff.			significance	separation
	bg = 1 %	bg = 10 %	bg = 30 %		
RE.	0.716	0.938	0.982	2.595	0.772
Boosted DT	0.673	0.931	0.978	2.436	0.752
Fisher	0.694	0.921	0.954	1.599	0.668
Neural Network	0.704	0.921	0.956	2.367	0.746
Likelihood	0.662	0.920	0.974	2.067	0.721
HMatrix	0.685	0.918	0.959	2.015	0.734

Table 6: Comparison of various multivariate methods. For a definition of the different figures of merit we refer to the text.

using this new weight. The procedure is repeated many times and in this way an ensemble of decision trees is built. The final score of the event will be determined by following each event through each tree in turn and build an (possibly weighted) average over all the trees in the ensemble. There are a number of different algorithms used to define the new weights of the misclassified events. The particular algorithm that we compare with is called *AdaBoost*. We refer the reader to [22] or to [4] for more details on this algorithm.

5.2 Results of the comparison

In table 6 we show the signal efficiency of the different methods at three different background efficiencies as well as the separation, defined as:

$$\frac{1}{2}\int_{-\infty}^{\infty}\frac{(S(x)-B(x))^2}{S(x)+B(x)}dx \tag{5.7}$$

where x is the obtained scores, and $S(x)$,$B(x)$ are the signal and background distributions of the scores, respectively. We also show the significance, defined as

$$\frac{(\bar{x}_S-\bar{x}_B)}{(rms_S)^2+(rms_B)^2} \tag{5.8}$$

where $\bar{x}_{S,B}$ and $\bar{x}_B$ are the mean score values found by MVA for signal and background, respectively and $rms_{S,B}$ are the corresponding root mean square values. In figure 7 the signal efficiency as a function of background efficiency is displayed for the different methods. Here only the four most important discriminating variables (as defined by *RE*) were used. From table 6 and figure 7 it can be concluded that *RE* performs best for all signal efficiencies and background efficiencies. Most pronounced is the difference, if one requires low background efficiencies. For high signal efficiencies boosted decision trees and ANN perform almost as good as *RE*, whereas their performance degrades significantly for low background efficiencies. Overall, ANN come closest in performance to *RE*.

In this comparison we choose a sample using only the four most relevant variables. It should be noted, that ANN (and probably also the other methods) in general show degraded performance if irrelevant discriminating variables are added to the sample.

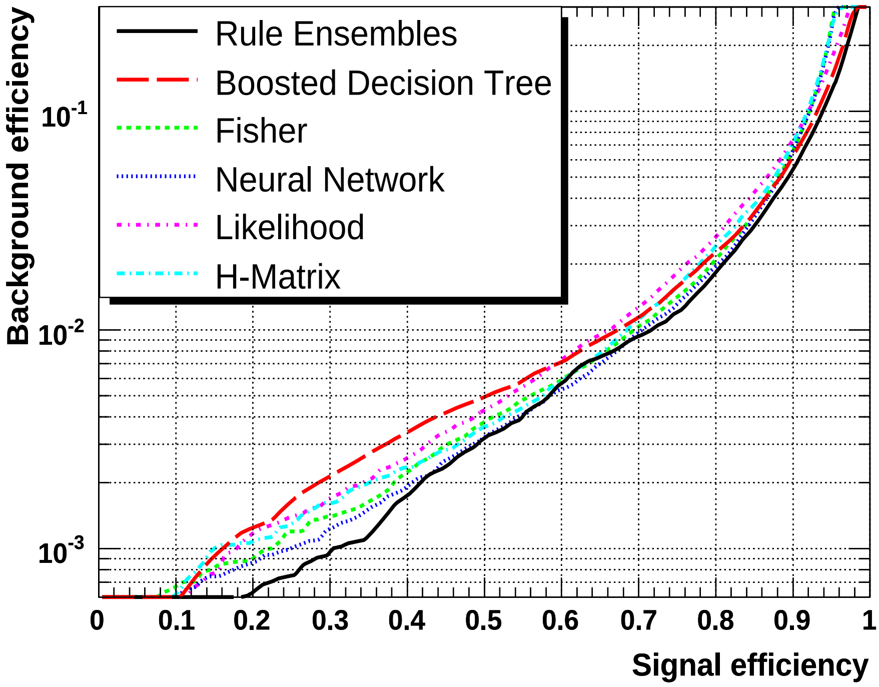

Figure 7: Comparison of efficiencies between various multi variate methods. Only the four best variables where used.

6. Discussion and conclusions

In this note we illustrate the use of a novel, powerful multi-variate classification method, called *RE* to reject top background in search for super-symmetric particles in ATLAS. The method is based on the concept of ensembles of learning machines. In particular, an ensemble of linear combinations of cuts in discriminant variable space is constructed by means of decision trees. Since linear combinations of cuts are used, *RE* are easy to diagnose: important cuts and discriminant variables are easily identified, which can be used to simplify the model or to decide which variable deserves extra attention. In our illustration we reduce the initial model of fourteen discriminant variables to a model only including the four most important without a major degradation of the performance of the method.

Since super-symmetry is so poorly constraint, training samples have to be constructed for benchmark points in the space of possible parameters. We consider in this paper two benchmark models within SUGRA. It is shown that the method is robust under training on the "wrong" signal model, indicating stability under systematic shifts of the discriminating variables. Training on the "wrong" model degrades the performance moderately. Using a mixture of different benchmark models for training is shown to be not efficient.

The top background is most probably the dominant background in the search for SUSY since its signature is hardest to discriminate from SUSY events. The presented method shows better discrimination power than the most common multi-variate methods used in high energy physics. The importance of choosing such a method should not be underestimated: the ratio of the amount of data which needs to be taken for a given significance scales like the square of the ratio of signal efficiencies for the two methods. In our example, the amount of data taken using RE for an expected background efficiency of 1 % is only 85 % of the amount of data which would be necessary if using the likelihood method.

Acknowledgments

We have made use of the ATLAS physics analysis framework and data samples which are the result of collaboration-wide efforts. The authors would like to thank Jerome Friedman for discussions. We would also like to thank the TMVA development team for their help.

References

[1] H.B. Prosper, *Multivariate methods: A unified perspective*, prepared for Conference on Advanced Statistical Techniques in Particle Physics, Durham, England, 18-22 Mar 2002.

[2] J.H. Friedman, B.E. Popescu, *Predictive learning via rule ensembles*, Stanford University, Department of Statistics, Technical Report, 2005.

[3] J.H. Friedman, *Statistical problems in particle physics, astrophysics and cosmology*, Oxford, England, United Kingdom, 12-15 Sep 2005, to appear in the proceedings of PHYSTATO5.

[4] B.P. Roe, H.J. Yang, J. Zhu, Y. Liu, I. Stancu and G. McGregor, *Boosted decision trees, an alternative to artificial neural networks*, *Nucl. Instrum. Meth. A* **543** (2005) 577 [physics/0408124].

[5] A.J. Smola, B. Schölkopf, NeuroCOLT2 Technical Report Series, NC2-TR-1998-030, 1998.

[6] J.H. Friedman, B.E. Popescu, *Importance samples learning ensembles*, Stanford University, Department of Statistics, Technical Report, 2003.

[7] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and regression trees*, Wadsworth International Group, Belmont, California, USA, 1983.

[8] J.H. Friedman, B.E. Popescu, *Gradient directed regularization for linear regression and classification*, Stanford University, Department of Statistics. Technical Report, 2004.

[9] *The R project for statistical computing*, code and documentation available from http://www.r-project.org/.

[10] See e.g. http://tmva.sourceforge.net/.

[11] F. Cooper, A. Khare and U. Sukhatme, *Supersymmetry and quantum mechanics*, *Phys. Rept.* **251** (1995) 267 [hep-th/9405029].

[12] ATLAS collaboration, W.W. Armstrong et al., *ATLAS: Technical proposal for a general-purpose p p experiment at the Large Hadron Collider at CERN*, CERN-LHCC-94-43.

[13] CMS collaboration, *CMS, the compact muon solenoid: technical proposal*, CERN-LHCC-94-38.

[14] LHCb collaboration, S. Amato et al. *LHCb technical proposal*, CERN-LHCC-98-4.

[15] ALICE collaboration, *ALICE: Technical proposal for a large ion collider experiment at the CERN LHC*, CERN-LHCC-95-71.

[16] GEANT4 collaboration, S. Agostinelli et al., *GEANT4: A simulation toolkit*, *Nucl. Instrum. Meth. A* **506** (2003) 250.

[17] ATLAS collaboration, *Atlas detector and physics performance technical design report*, CERN/LHCC 99-15, 1999.

[18] H.P. Nilles, *Phys Reports* **110**.

[19] Particle Data Group collaboration, S. Eidelman et al., *Review of particle physics*, *Phys. Lett.* **B 592** (2004) 1.

[20] S. Weinberg, *Implications of dynamical symmetry breaking*, *Phys. Rev.* **D 13** (1976) 974; *Implications of dynamical symmetry breaking: an addendum*, *Phys. Rev.* **D 19** (1979) 1277.

[21] O. Bruning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole and P Proudlock, *LHC design report, vol. I. The LHC main ring*, CERN-2004-003.

[22] Y. Freund and R.E. Schapire, Proc. COLT, 209-217, ACM Press, New York, 1996.