Statistical classification techniques in high energy physics (SDDT algorithm)

Petr Bouř, Václav Kůs and Jiří Franc
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Trojanova 13, 12000 Praha 2, Czech Republic
E-mail: petr.bour@fjfi.cvut.cz, vaclav.kus@fjfi.cvut.cz, jiri.franc@fjfi.cvut.cz

Abstract. We present our proposal of the supervised binary divergence decision tree with nested separation method based on the generalized linear models. A key insight we provide is the clustering driven only by a few selected physical variables. The proper selection consists of the variables achieving the maximal divergence measure between two different classes. Further, we apply our method to Monte Carlo simulations of physics processes corresponding to a data sample of top quark-antiquark pair candidate events in the lepton+jets decay channel. The data sample is produced in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV. It corresponds to an integrated luminosity of 9.7 $fb^{-1}$ recorded with the D0 detector during Run II of the Fermilab Tevatron Collider. The efficiency of our algorithm achieves 90% AUC in separating signal from background. We also briefly deal with the modification of statistical tests applicable to weighted data sets in order to test homogeneity of the Monte Carlo simulations and measured data. The justification of these modified tests is proposed through the divergence tests.

1. Introduction and Data Sample
Experiments running on particle accelerators produce a vast amount of raw data containing not only the crucial signal but also a large number of observations originating from another physical processes (backgrounds). In this paper, we are concerned with signal separation from massively contaminated data measured with the D0 detector during Run II of the Fermilab Tevatron Collider [1]. We are interested in top-antitop quark pair production after a collision of accelerated protons and antiprotons with one lepton (electron or muon) and at least two jets in the final state. This study is based upon the full Tevatron data sample recorded by the D0 detector at $\sqrt{s} = 1.96$ TeV and, after applying data quality requirements, corresponding to an integrated luminosity of 9.7 $fb^{-1}$ [2]. The selection in the lepton+jets decay channel is similar to the one described in [3].

Let us denote the particular class of observations by $S$ (signal) and the class of different decay processes by $B$ (backgrounds). In order to train our classification algorithm, we are provided with the Monte Carlo (MC) simulations coming from different MC event generators, but with known classification to $S$ or $B$. The dimension of MC is limited to $m = 46$ selected physical variables. After validation of homogeneity between data and MC, we perform supervised classification. However, there is a weighting applied to MC we have to cope with.
2. Divergence Measures and Variable Selection

Due to the nature of our separation techniques we use the specific divergence measures as a tool for wise variable selection in every step of our signal separation. Since some variables are strongly correlated with each other and they differ in signal separability, we select only those variables of the highest distinction between probability distributions of $S$ and $B$. For this purpose, we define normalized Rényi distance of order $\alpha$ for distributions $P, Q$ with probability densities $p, q$ as $R_\alpha(P, Q) = \frac{1}{\alpha(1-\alpha)} \int p^\alpha q^{1-\alpha} \, d\mu$, where $\alpha \in \mathbb{R}$, $\alpha \neq 0$, $\alpha \neq 1$, and $\mu$ is a dominating measure on $\mathbb{R}$. Moreover, by [4], we have $R_1(P, Q) = \lim_{\alpha \rightarrow 1} R_\alpha(P, Q)$, and $R_0(P, Q) = \lim_{\alpha \rightarrow 0} R_\alpha(P, Q)$, to be the reversed Kullback-Leibler divergence $K L(Q, P)$, and $K L(P, Q)$, respectively. It is easily seen that it holds $R_\alpha(P, Q) = R_{1-\alpha}(Q, P)$.

To evaluate the $R_\alpha(P, Q)$ divergence for weighted MC, we consider the modified weighted empirical distribution function (EDF) based on an analogy with EDF for weighted inputs. Let $X = (X_1, \ldots, X_n)$ be iid random variables distributed by cumulative distribution function (CDF) $F(x)$ and let $(w_1, \ldots, w_n)$ be respective weights, where $W = \sum_{i=1}^n w_i$. We define the weighted empirical distribution function (WEDF) to be $F_n^W(x) = (1/W) \sum_{i=1}^n w_i I_{(-\infty,x]}(X_i)$, where $I_A(X)$ is the indicator of the set $A$. In the case of $w_i = 1$ for all $i \in \mathbb{n}$, that is the unweighted recorded data, the definition of WEDF coincides with the usual EDF. The continuous version of WEDF serve for finding of the weighted histograms with equiprobable binnings, see [5] in detail. This continuous estimate of EDF is essential for robust equiprobable binning because the value of the normalized Rényi distance is independent on the bin width and also on the number of bins, provided we keep the number of observations in a bin at a reasonable level. Equiprobable binning with corresponding histograms of $S$ and $B$ classes for selected variables in MC can be seen in figure 1. Notice the considerable similarity of both distributions.

![Figure 1](image1.png)

**Figure 1.** Weighted histograms with equiprobable binnings for variables $\Delta \phi (j^1, j^2)$, $\eta^{j_1}$ and $K_t^{m_{j_1}}$ in MC channel Electron 2 Jets.

During our classification algorithm we need to select the variables potentially useful for separation of $S$ and $B$. Therefore, we are looking for those variables with the highest value of divergence $R_\alpha(P, Q)$ between $S$ and $B$. The question naturally arises whether we shall use the distribution of $S$ or $B$ in the first argument of $R_\alpha(P, Q)$. Which value of the parameter $\alpha$ should be chosen is also of some interest. Fortunately, there is a very satisfactory and consistent solution to this problem, due to the specific symmetry mentioned above. Observe the dependence of the symmetry with its divergence values on parameter $\alpha$ in figure 2. Thus we choose $\alpha = \frac{1}{2}$, which provides us with the unique fixed point for comparison of all variables since $R_{1/2}(S, B) = R_{1/2}(B, S)$.

3. Supervised Divergence Decision Tree

The binary supervised divergence decision tree (SDDT) was originally proposed and tested in [5]. This universal classification method is based upon certain localization of separation
among nodes of the tree. We shall briefly describe the final algorithm. Suppose we are given weighted simulation MC ⊂ \mathbb{R}^m, where m stands for the number of MC variables. For every observation \( x_i \in \text{MC} \) there is a weight \( w_i > 0 \). Denote by \( X, Y \subset \text{MC} \) the training set and test set, respectively. Recall the definition of the signal class \( S \) and the backgrounds class \( B \) from section 1. Let \( U = \{\{U_X\}, \{U_Y\}\} \) be a current node for the algorithm, where \( U_X \subset X, U_Y \subset Y \). After the separation of the node \( U \), we denote by \( U_0, U_1 \subset U \) the sets of observations classified as \( B, S \), respectively. For the simplicity, we write \( U^{(k)} \) for the current node, where \( k \) is an unique binary code reflecting the path in the tree. Let \( M^{(k)} \) be the separation model produced in the node \( U^{(k)} \). For every \( x_i \in U^{(k)} \) we obtain the probability \( p_i^{(k)} \) of signal classification by the model \( M^{(k)} \). Next, we insert \( X, Y \) into the root of the tree and declare the root as a current node \( U = \{\{U_X\}, \{U_Y\}\} = \{\{X\}, \{Y\}\} \). From now on, the SDDT learning algorithm proceeds as follows:

(i) For appropriately chosen \( \gamma > 0 \), if \( \sum_{x_i \in U_X \cap S} w_i < \gamma \) or \( \sum_{x_i \in U_X \cap B} w_i < \gamma \) then we declare the current node \( U \) as a pure final leaf.

(ii) For appropriately chosen \( d \in \hat{m} \) and for every combination of \( d \)-dimensional selections from \( m \) variables, under use of histogram with equiprobable binning, we compute the \( \phi \)-divergence \( D_{\phi}^{(j)}(S, B) \) in the current node \( U \), where \( j \in \{1, \ldots, \binom{m}{d}\} \). We used here the normalized Rényi distance of order \( \alpha \).

(iii) From all of the \( \binom{m}{d} \) combinations we select the \( j^* \)-th combination such that the \( \phi \)-divergence \( D_{\phi}^{(j^*)}(S, B) \) is maximal in \( U \), i.e. \( j^* = \arg \max \{D_{\phi}^{(j)}(S, B) : j = 1, \ldots, \binom{m}{d}\} \).

(iv) Only for \( d \) selected variables in \( j^* \)-th combination we are training the nested classification based on generalized linear models ([6]) on the training set \( U_X \). Further, we optimize and test the produced model \( M^{(k)} \) on the test set \( U_Y \).

(v) In the case where overfitting is detected, we repeat step iv with the different setting of parameters in the classification tool. Otherwise, we record the model \( M^{(k)} \) and the probability \( p_i^{(k)} \) for every \( x_i \in U \).

(vi) We split the node \( U \) with the model \( M^{(k)} \) per \( d \) selected variables in \( j^* \)-th combination into two new nodes \( U_0, U_1 \).

Following the algorithm we arrive at probability vector \( D_i = (p_i^{(k_1)}, \ldots, p_i^{(k_t)}) \) for each \( x_i \in \{X \cup Y\} \), where \( l_i \) is the number of levels in the tree for \( x_i \). The resulting discriminant of probabilities of signal classification \( D = (p_1, \ldots, p_N) \) can be obtained in several manners indicated in [5], including for example processing the discriminants \( D_i \) by artificial neural

![Figure 2. Dependence of normalized Rényi distance on the parameter α for distributions of S and B for variables Δφ(j1,j2), ηj1 and Kminp in MC channel Electron 2 Jets.](image)
networks or even by the secondary application of SDDT to discriminants $D_i$. For the results in section 4 we take into account the arithmetic mean of the discriminant $D_i$.

![Figure 3. ROC for SDDT signal separation in MC Electron 4+ Jets and Muon 4+ Jets channels.](image)

**Table 1.** AUC for SDDT signal separation in MC Electron 4+ Jets and Muon 4+ Jets channels.

|   | Electron 4+ Jets | Muon 4+ Jets |
|---|-----------------|--------------|
| 2 | 0.8840 training | 0.8895 training |
|   | 0.8694 test     | 0.8900 test  |
|   | 0.8764 yield    | 0.8884 yield |
| 3 | 0.8863 training | 0.8908 training |
|   | 0.8710 test     | 0.8909 test  |
|   | 0.8776 yield    | 0.8901 yield |
| 4 | 0.8876 training | 0.8942 training |
|   | 0.8731 test     | 0.8950 test  |
|   | 0.8789 yield    | 0.8935 yield |

![Figure 4. Optimal cut $\delta^* = 0.284$](image)

**4. Study of signal separation in MC $t\bar{t}$ lepton+jets decay channel**

Now let us describe the SDDT separation of signal $S$ from MC D0. Prior to MC classification, we tested the performance of SDDT in order to reflect the dependence on the choice of parameters $\gamma, d$ and $\alpha$. The success of SDDT separation is measured via the Area Under Curve (AUC)
of the Receiver Operating Characteristic (ROC), as observed in figure 3. Generally, AUC for SDDT separation grows with higher dimension $d$ and remains without notable overfitting. As an example, see table 1, where AUC results are quite comparable with AUC obtained in [7]. We close this section by showing the example of the optimal cut $\delta^*$ such that the observations $x_i$ are classified as signal $S$ iff $p_i \geq \delta^*$. For this cutting point we can see the overall error of classification ERR and the error $ERR_s$ within signals $S$ in figure 4. The universality of the originally proposed SDDT method consists in almost arbitrary choice of nested separation technique. Among others, there is a great potential in optimization of the whole SDDT procedure through the appropriate choice of SDDT parameters $\gamma, d, \alpha$. However, the profound research still needs to be carried out.

5. Addendum: Weighted Tests of Homogeneity

Prior to supervised learning of any classification method, it is vital to guarantee homogeneity of measured data and MC distributions. In order to avoid the robustness problems of a particular parametric family (for the normal distribution case see [8]), we pursued nonparametric homogeneity testing based upon EDFs of two data sets $X = (X_1, \ldots, X_{n_1})$, $Y = (Y_1, \ldots, Y_{n_2})$, with respective distribution functions $F, G$. By the homogeneity hypothesis, as our null hypothesis $H_0$, we understand $H_0 : F = G$ vs $H_1 : F \neq G$ at significance level $\alpha \in (0, 1)$. The nature of homogeneity testing prompted us to look for the $p$-value, i.e., the lowest significance level for which we reject the hypothesis $H_0$. We considered two homogeneity tests based on EDFs of our data sets: two sample Kolmogorov-Smirnov test [9], and two sample Anderson-Darling test [10]. However, for weighted data samples we were forced to replace EDFs with their respective WEDFs. These modified tests for the weighted data samples do not have to obey the asymptotic statistical properties of the original procedures. Let us stress that the $p$-value obtained using the modified weighted tests can not be considered a regular approximate $p$-value without subsequent detailed research. However, we carried out the numerical verification of the modified approaches in [5] and found there is no significant difference observed in $p$-values related to regular unweighted tests.

Acknowledgments

This work was supported by the MEYS CZ grants LG15047, LM2015068, SGS15/214/OHK4/3T/14, and GA16-09848S.

References

[1] D0 Collaboration Abazov V M et al. 2006 The upgraded D0 detector Nucl.Instrum.Methods Sect.A 565, 463
[2] Casey B C K et al. 2013 The D0 Run Iib Luminosity Measurement Nucl. Instrum. Methods Sect. A 698, 208
[3] D0 Collaboration Abazov V M et al. 2014 Measurement of differential $tt$ production cross sections in $pp$ collisions Phys. Rev. D 90, 092006
[4] Vajda I 2006 On divergences and informations in statistics and information theory IEEE Trans. Inf. Theory 52 4394–442
[5] Bour P 2015 Testing of separation approaches with attributes selections for physical data (FNAL/CERN) (Prague: SPMS 2015 proceedings, FNSPE CTU in Prague)
[6] McCulloch C E, Searle S R and Neuhaus J M 2008 Generalized, Linear, and Mixed Models (New York: Wiley)
[7] Stepanek M, Franc J and Kus V 2015 Modification of Gaussian mixture models for data classification in high energy physics Journal of Physics: Conference Series 574 012150
[8] Frydlova I, Vajda I and Kus V 2012 Modified power divergence estimators in normal model- simulation and comparative study, Kybernetika 48(4) 795–808
[9] Smirnov N J 1944 Approximate laws of distribution of random variables from empirical data Usp. Mat. Nauk 10 179–206
[10] Engmann S and Cousineau D 2011 Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test J. Appl. Quant. Methods 6 1–17