Volatility of selected separators/classifiers wrt. data sets from field of particle physics

Jiřina, Marcel
2011

Institute of Computer Science
Academy of Sciences of the Czech Republic

Volatility of selected separators/classifiers wrt. data sets from field of particle physics

Marcel Jiřina and František Hakl

Technical Report No. V-1126

November 2011

Abstract

We study the volatility, i.e. influence of random changes in data sets to overall separation/classification behavior of separators/classifiers. This is motivated by the fact, that simulated data and true data from ATLAS experiment may differ, and a question arises what if separators or cuts are optimized for simulated data, and then used for true data from the experiment. This behavior was studied using simulated data modified by artificial distortions of known size. We found that even slight change in data sets causes a little worse result than supposed but, surprisingly, even relatively large distortions give then nearly the same results. Only truly great variations cause degradation of separation quality of separator/classifier as well as of the cuts method.

Keywords:

Multivariate data, volatility, classification, signal-background separation, physics event data, particle physics.

Volatility of selected separators/classifiers wrt. data sets from field of particle physics

Marcel Jiřina and František Hakl

Contents

Introduction

This work is motivated by the fact, that simulated data and true data from ATLAS experiment for exactly the same task or problem may differ. Than a question arises what if separators or cuts are optimized for simulated data, and then used for true data from the experiment.

The changes caused by such a difference in data we call a volatility here. We could also use term“ sensitivity”, but this term is already reserved to a different thing, to the ability of a separation or a classification method to keep useful data usually called a signal as much as possible. The sensitivity is thus the same as ”signal acceptance” or “signal efficiency” and is depicted in a ROC graph on a vertical axis. The term volatility we borrowed from econometrics where this notion describes the changes on a market, especially the market of shares. When there is a low volatility, the market (in prices or volumes) changes slightly or in a steady way, a high or a large volatility means an unstable market with large changes up and down. In our use here analogically a low volatility means small changes in separator/classifier -data characteristics, high volatility means large changes in these characteristics.

We found that for rather moderate differences in data mentioned above there is a low volatility of classifiers as well as in CUTs method. It means that even slight change in data sets causes a little worse result, compared to original ones, but, surprisingly, even relatively large distortions give nearly the same results. Making variations in data larger we found that only truly great variations cause degradation of separation quality of a separator/classifier as well as of the cuts method. Thus the message of this study is that in any case results with true data will be necessarily a little worse than for simulated data, but the change as large as ten per cent in individual variables causes the same change as 0.001 per cent changes.

Data sets description

Seven variables data set “Elsbieta 7”

Identification of hadronic decays will be the key to the possible Higgs boson discovery in the wide range of the MSSM parameter space. The $h/H/A \rightarrow \tau\tau$ and $H^{\pm} \rightarrow \tau\upsilon$ are promising channels in the mass range spanning from roughly 100 GeV to 800 GeV. The sensitivity increases with large $\tan\beta$ and decreases with rising mass of the Higgs boson. The $H \rightarrow \tau\tau$ decays will give access to the Standard Model and light Minimal Supersymmetric Standard Model Higgs boson observability around m_H = 120 GeV, with Higgs boson produced by vector-boson fusion. The hadronic τ identification is also very important in searching for supersymmetric particles, particularly at high $\tan\beta$ values.

In this data as signal, we consider reconstructed candidates from tau decays in $pp \rightarrow W \rightarrow \tau\nu$ and $pp \rightarrow Z \rightarrow \tau\tau$ events. As background, we consider candidates from QCD shower in the same $pp \rightarrow W \rightarrow \tau\nu$, $pp \rightarrow Z \rightarrow \tau\tau$ events and in QCD dijet events (sample with $p_T^{\text{hard}} >$ 35 GeV).

In our test we used data tau-3Pwtoenu-0-200-GeV-lrn.dta and tau-3Pwtoenu-0-200-GeV-tst.dta having 7 variables. We do not deal with them in detail here as it may be found in [2] and we reproduce it verbatim in Table 1. This data uses three-prong candidates that are seeded by the bary-center of three nearby tracks. At the same time, full scale from zero to 200 GeV Higgs boson mass is used, i.e. no cuts are used.

Table 1. Data set description

For the classification procedure calorimetric observables as described in details in [4] are used. Separately we optimize identification procedure for single-prong (τ_{1P}) and three-prong (τ_{3P}) candidates. The τ_{1P} is seeded by the leading hadronic track at vertex (track η and ϕ at the vertex). The τ_{3P} is seeded by the bary-center of three nearby tracks. The calorimetric observables are calculated from energy deposition in cells within a distance from a seed of $\Delta R = 0.2$.

The following calorimetric and tracking variables are used to build discriminating observables:

- Track transverse momenta of a leading track p_T^{track} (or scalar sum of tracks transverse momenta in case of τ_{3P} candidates)
- Electromagnetic radius of the τ-candidate, R_{em}
- Number of strips N_{strips}^{τ} , strips with energy deposition above a certain threshold
- The width of energy deposition in strips, W_{strips}^{τ}
- The fraction of the transverse energy deposited, $fracET_{R12}$, in the $0.1 < \Delta R < 0.2$ radius with respect to the total energy in the cone $\Delta R = 0.2$. Cells belonging to all layers of the calorimeter are used.
- The ratio of energy deposited in the hadronic calorimeter $E_T^{chrgHAD}$ and track transverse momenta, $\frac{E_T^{chrgHAD}}{p_T^{track}}$ (or sum of transverse momenta in case of τ_{3P} candidates)
- The ratio of energy deposited in calorimeters in a a ring $0.2 < \Delta R < 0.4$, with respect to the total energy deposited in a cone $\Delta R < 0.4$, E_T^{chrgEM}/E_T^{calo} and $E_T^{chrgHAD}/E_T^{calo}$.

The variables above are used either directly or to build up in total 6 discriminating variables: N_{strips}^{τ}, W_{strips}^{τ}, $fracET_{R12}$, R_{em}, $\frac{E_T^{chrgHAD}}{p_T^{track}}$, $\frac{E_T^{otherEM}+E_T^{otherHAD}}{E_T^{calo}}$. Classifiers use them separately, without any assumptions on the possible correlations.

Data sets modification

Each variable v_i, $i = 1,...7$ of the original data sets has been perturbated by adding a random errors with normal distribution density. These errors has been produced to have a zero mean value and variance equal to mean value of the original variable (= $\mu(v_i)$) multiplied by a volatility parameters (= v_p); we use this volatility parameters set to 10^{-6}, 10^{-5}, ... , 0.1+0.05*k, k=0,...8, 1.0. In other words, we substitute each original variable v_i by the new one, $v'_i = v_i + N(0,\mu(v_i)*v_p$.

Classifiers/separators used

To make terminology clear, we use word classifier for tool that is able to recognize samples, i.e. events of two or more kinds, classes. Separators discriminate between two classes only. For our needs all devices work as separators as we have two classes, signal and background only. Generally we can speak about classifiers.

In this study we used IINC classifier/separator, NNSU, the Neural Network with Switching Units, and standard cuts method for this data as described in [2]

IINC classifier

IINC is the Inverse Indexes of Neighbors Classifier [3], [4]. This relatively simple method was derived on the bases of estimating multifractal dimensions (Hurst exponents) and Zipfian distribution. Here we use it with L_1 (Manhattan) metrics, as we found generally better

behavior than with L_2 (Euclidean) metrics. The software is available for noncommercial use at http://www.marceljirina.cz/index.php?s=software&a=IINC0100 and can be run on Windows as well as on LINUX machines.

NNSU separator

NNSU (Neural Network with Switching Units) is separator based on genetic optimization of the general topology of neural networks. In addition, instead of classical neuronal units (like in multilayer perceptron model) it exploits switching units dividing feature space into disjoint subsets. We showed and broadly proved it's convenient to HEP data separation (e.g. in [2]). Distributed implementation of this separation tool is available on the site http://www.cs.cas.cz/nnsu/ for all expert community.

CUTs method

In this study we used the same cuts method as in [2]; verbatim description follows in Table 2 below.

Table 2. CUTs method description.

The cuts-based approach uses a sequence of properly tuned cuts for individual variables. The cuts used here for reference selection are as follows:

- $N^{\tau}_{strips} < 15$;
- $W^{\tau}_{strips} < 0.004$;
- $fracET_{R12} < 0.4$ for τ_{1P} (< 0.6 for τ_{3P});
- $R_{em} < 0.08$;
- $\frac{E_T^{chrgHAD}}{p_T^{track}} < 1.0$;
- $\frac{E_T^{otherEM}+E_T^{otherHAD}}{E_T^{calo}} < 0.15$ for τ_{1P} (< 0.25 for τ_{3P}).

It is obvious that the order of cuts in the sequence has no impact on the final acceptance.

Results

For this seven variables data set a typical ROC curve is depicted in Fig. 1 and 2.

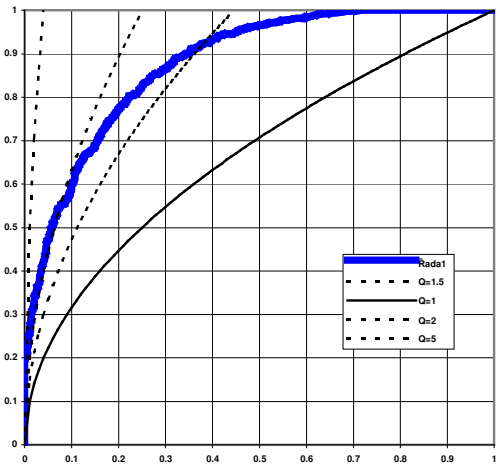

Fig. 1. ROC curve for data "Elsbieta 7", smoothed.

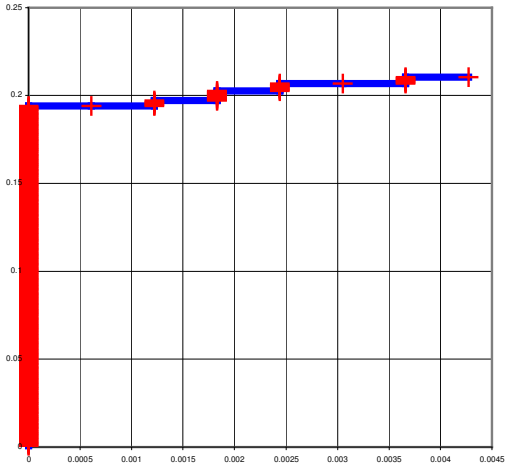

Fig. 2. ROC curve for data "Elsbieta 7", not smoothed, the left end detail. Red crosses indicate individual events.

In this study we modified the testing set according to description above. We also tried to modify the learning set the same way to show that there is no practical difference whether the difference is made in the testing or in the learning set.

In Table 3 and Figs. 3 and 4 an influence of perturbation size (the variation) in data to the minimal classification error. In Fig 3 it is seen that for small variations the minimal classification error is nearly constant. Fig. 4 shows that starting with variation 0.1, i.e. 10 % the error grows practically linearly with variation. It is also seen that there is no important difference between variation in the testing set and in the learning set. Based on this finding, tests with NNSU were limited to variation in the testing set only.

Table 3. Minimal classification errors for data with Gaussian noise added. IINC classifier.

Variation	0	1E-06	1E-05	0.0001	0.001	0.01	0.1	0.15
in LRN	18.97%	20.48%	20.48%	20.48%	20.48%	20.30%	20.77%	21.17%
in TST	18.79%	20.64%	20.64%	20.64%	20.59%	20.55%	20.70%	21.09%
Variation	0.2	0.25	0.3	0.35	0.4	0.45	0.5	1
in LRN	22.12%	22.89%	23.73%	25.47%	25.92%	27.24%	27.13%	34.42%
in TST	21.68%	21.78%	22.95%	23.04%	23.28%	24.42%	24.85%	29.07%

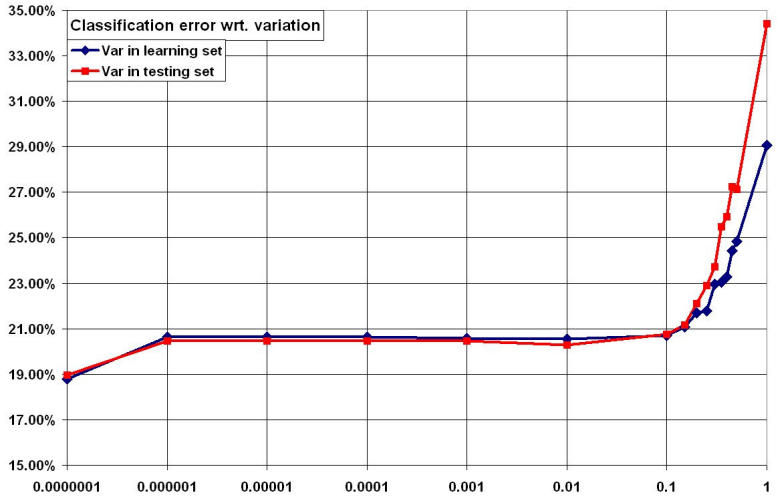

Fig. 3. The minimal classification error as a function of variation in logarithmic scale. IINC classifier.

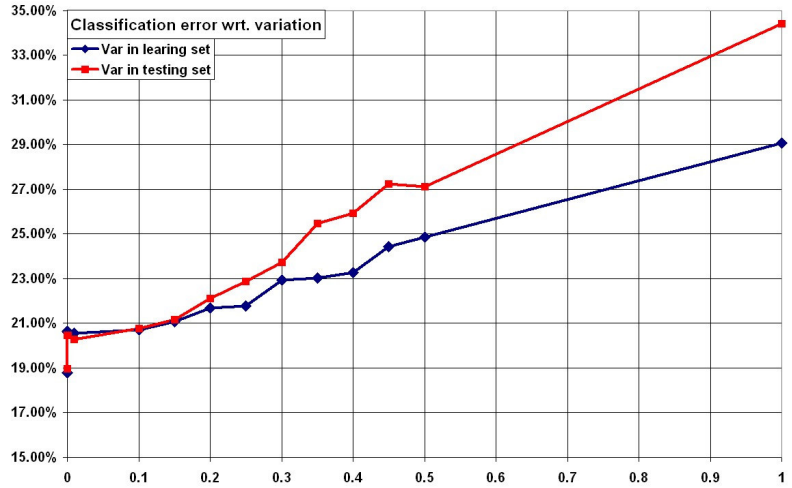

Fig. 4. The minimal classification error as a function of variation in linear scale. IINC classifier.

Table 4 and Figs. 5 and 6 show influence of perturbation size under the same conditions as above for the NNSU separator. One can see nearly identical results and pictures as above, i.e. for IINC classifier.

Table 4. Minimal classification errors for data with Gaussian noise added. IINC classifier.

Variation	1E-07	1E-06	1E-05	0.0001	0.001	0.01	0.1	0.15
NNSU	19.97%	20.68%	20.69%	20.68%	20.68%	20.67%	20.88%	20.88%
Variation	0.2	0.25	0.3	0.35	0.4	0.45	0.5	1
NNSU	21.83%	22.75%	23.53%	25.21%	26.13%	27.56%	27.45%	35.30%

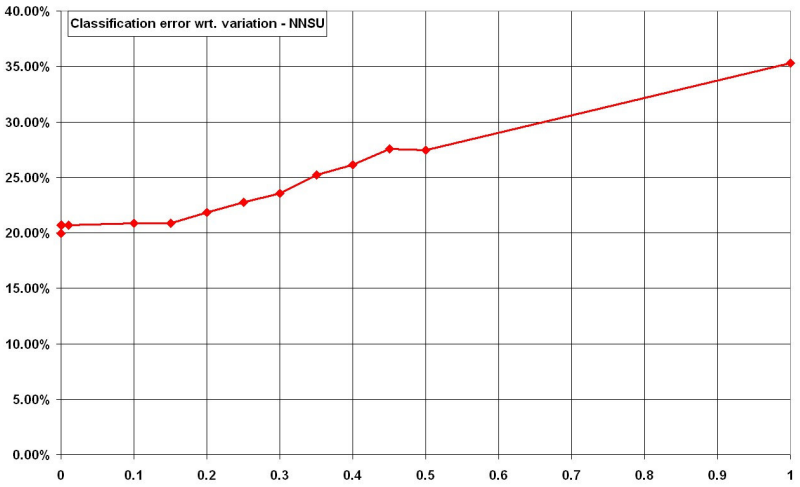

Fig. 5. The minimal classification error as a function of variation in linear scale. NNSU separator.

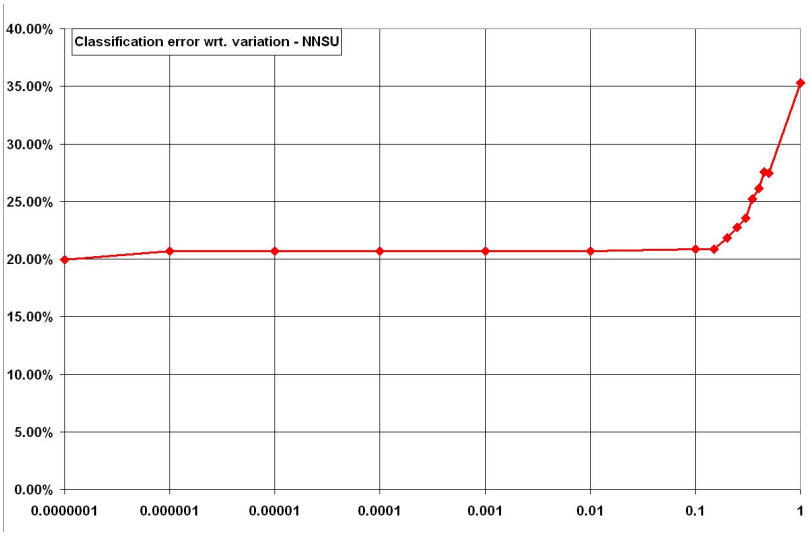

Fig. 6. The minimal classification error as a function of variation in linear scale. NNSU separator.

For the same data there are ROC curves shown. Again for IINC and NNSU separators/classifiers a different tint of red, orange or yellow and different tint of blue lines correspond to data with variation in the testing set and with variation in the learning set, respectively. To the uppermost (best) ROC curve corresponds the uppermost black diamond that denotes results obtained by the CUTs method. A group of lines, part of them marked by ellipsis in graph as well as in legend corresponds to variations between 0.000001 and 0.1 for variation in the learning set as well as in the testing set. To these cases a second diamond for CUTs method corresponds. The light blue and yellow ROC curves correspond to variations large as 1, i.e. 100 % in learning and in the testing set respectively. It is apparent that that are degenerated cases also corresponding to the lowermost diamond for the CUTs method

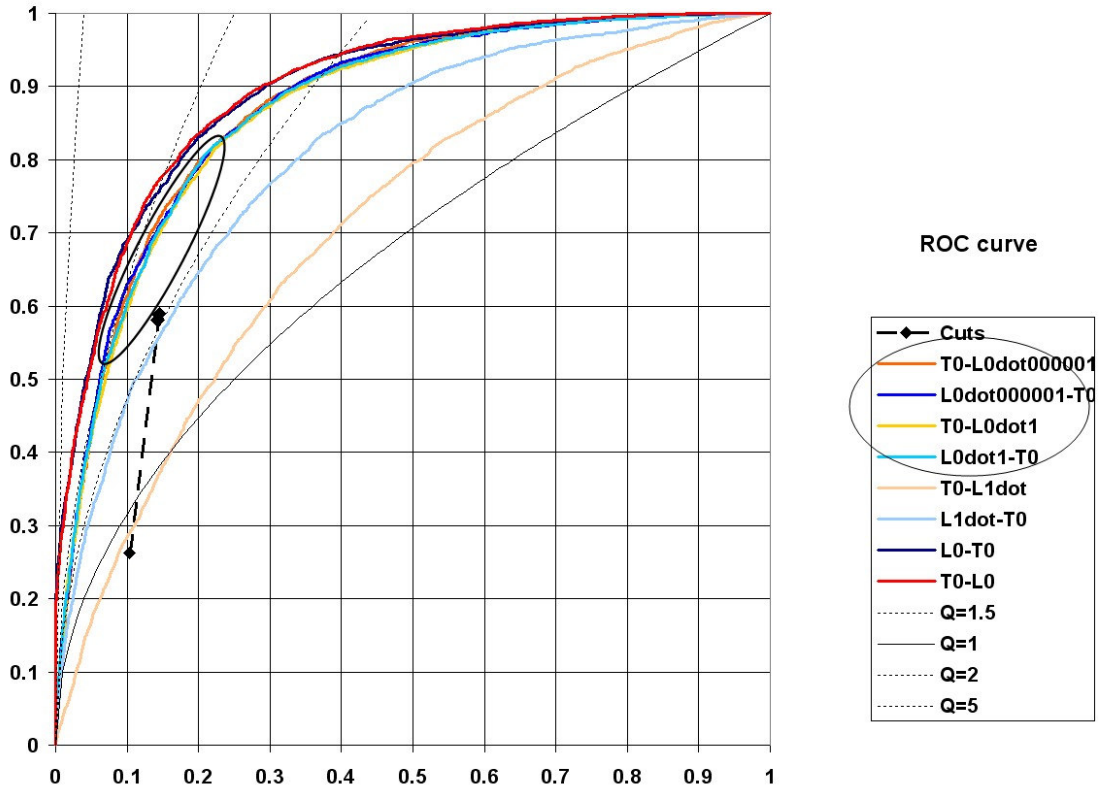

Fig. 7. The ROC curves for different values of variation. Note red and blue lines corresponding to L0-T0 and T0-L0 data, i.e. data without any variation. This case gives the best results and also corresponds to uppermost black diamond denoting results obtained by the CUTs method. Then note a group of lines, part of them marked by ellipsis in graph as well as in legend. These lines correspond to variations between 0.000001 and 0.1 for variation in the learning set as well as in the testing set. To these cases a second diamond for CUTs method corresponds. The light blue and yellow ROC curves correspond to variations large as 1, i.e. 100 % in learning and in the testing set respectively. It is apparent that that are degenerated cases also corresponding to the lowermost diamond for the CUTs method.

In Fig. 8 there are ROC curves for different variations of data in the testing set obtained for the NNSU separator. It is easily seen that picture is nearly the same as in Fig. 7.

Note that in both figures, Fig. 7 and Fig. 8 there are black dotted lines and a thin black line. These lines represent constant values of so-called quality factor $Q = S/\sqrt{B}$. In fact if data before separator has this ratio equal to Q_{in}, then data accepted as a signal has this ratio equal to $Q_{\text{out}} = Q.Q_{\text{in}}$. In cases depicted in Figs. 7 and 8 one can see that the best value of Q can be reached 2.2 and even for data with variation 10 % there is a region on the ROC curve with $Q = 2$. At the same time, with the CUTs method one can reach Q slightly above 1.5 only.

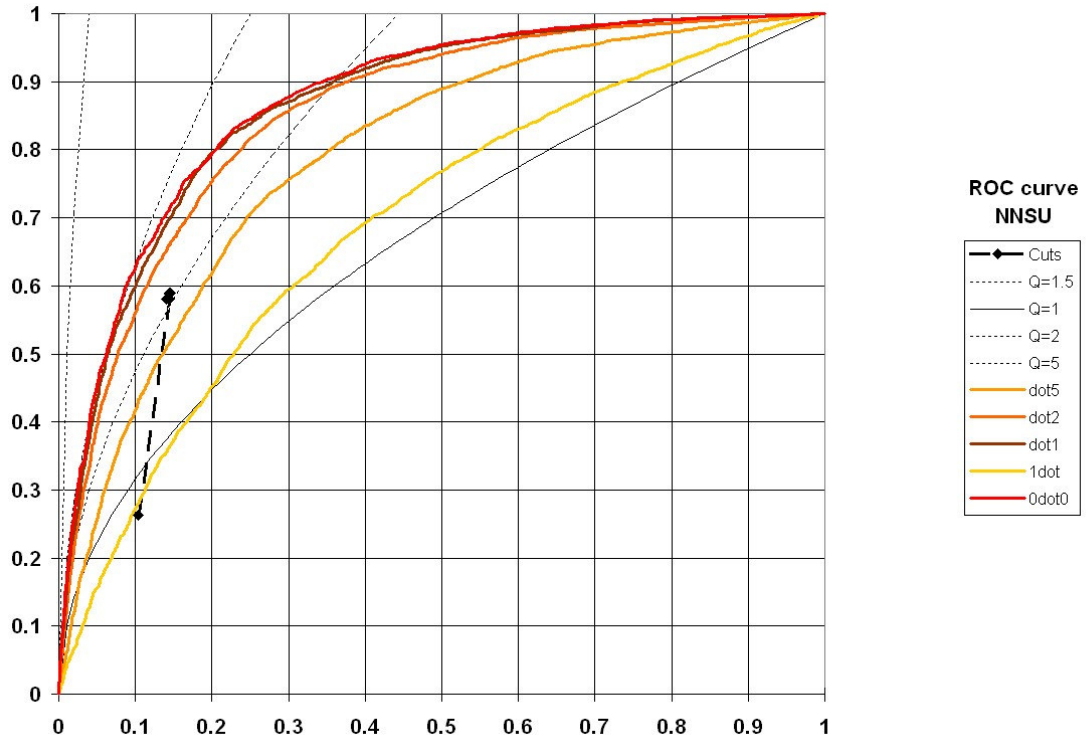

Fig. 8. The ROC curves for different values of variation from zero to 100 %. Red line 0dot0 correspond to data without any variation. This case gives the best results and also corresponds to uppermost black diamond denoting results obtained by the CUTs method. Brown line corresponds to variations between 0.000001 and 0.1 for variation in the testing set. To these cases a second diamond for CUTs method corresponds. Two gold and yellow lines correspond to variations 50 % and 100 % in the testing set respectively. It is apparent that that are degenerated cases also corresponding to the lowermost diamond for the CUTs method and that these results are very close to those obtained by IINC classifier.

Discussion

This study of the volatility, i.e. influence of random changes in data sets to overall separation/classification behavior of separators/classifiers is motivated by the fact, that simulated data and true data from ATLAS experiment may differ. We try to answer a question what happens if separators or cuts optimized for simulated data are subsequently used for the true data from the experiment.

We used simulated data and add some quantifiable and known amount of normal noise to all data variables. Dr. Elsbieta Richter-Was provided simulated data; data is the same as described and studied in several previous reports and in ATLAS note [2]

Our results can be summarized as follows:

- Even slight amount of noise (0.000001, i.e. 0.0001 per cent) in data sets causes a little worse result than without noise (20.64% vs. 18.79% for minimal separation/classification error of the IINC classifier that uses fractal nature of data, and 20.68% vs. 19.97% for NNSU that uses genetic optimization and clustering.

- This small difference remains nearly constant until 10 per cent noise; see the upper part of Table Z and Fig. A. It means that even relatively large distortions give nearly the same results in terms of separation quality.
- Then, i.e. for noise larger than 10 %, the minimal classification error grows nearly linearly with noise as depicted in Fig. B until large degradation for 100 % noise. Even for this big noise the classification error is equal to 30 or 34 percent for noise in the learning and testing set, respectively.

From these simple facts one can conclude that any small deviation from data used for setting the separating method (learning and testing or for setting cuts in CUTs method) causes minor degradation of results, but, surprisingly, even for relatively large (in the sense of 10 %) deviation in data this degradation remains the same. Only truly great variations cause degradation of separation quality of separator/classifier as well as of the cuts method.

We suppose that by application of noise with Gaussian distribution that has unlimited tails some artificial outliers may eventually appear. Outliers cause a small degradation of results as reported. We suppose that data processing procedures used before are designed so that absolute values of data items are limited and thus outliers are not contained in experimental data in this stage of processing.

We can conclude that separators used here and tuned according to simulated data are robust to relatively large differences between simulated and measured data from the ATLAS experiment.

Acknowledgement

This work was supported in part by the Institute of Physics of the Academy of Sciences of the Czech Republic under contract to ISC AS CR and in part by the Ministry of Education of the Czech Republic under the project Center of Applied Cybernetics No. 1M0567.

References

[1] Jiřina,M., Jiřina,M.,jr.: Testing Random Forest for Unix and Windows. Technical Report No. V-1075, Institue of Computer Science AS CR, Prague (2010)

[2] Hakl,F., Jiřina,M., Richter-Was,E.: Hadronic tau's identification using artificial neural network. ATLAS Physics Communication, ATL-COM-PHYS-2005-044, last revision: 26 August, 12 pp. (2005)

[3] M. Jiřina, M. Jiřina Jr.: Classifier Based on Inverted Indexes of Neighbors. Technical Report No. V-1034, Institute of Computer Science, Academy of Sciences of the Czech Republic, 11 pp., 2008.

[4] M. Jiřina, M. Jiřina Jr.: Classifier Based on Inverted Indexes of Neighbors II. - Theory and Appendix. Technical Report No. V-1041, Institute of Computer Science, Academy of Sciences of the Czech Republic, 26 pp., 2008.

***