



národní
úložiště
šedé
literatury

Testing selected separators/classifiers on simulated data sets from field of particle physics

Jiřina, Marcel
2010

Dostupný z <http://www.nusl.cz/ntk/nusl-41905>

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 10.04.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní nusl.cz.



Institute of Computer Science
Academy of Sciences of the Czech Republic

Testing selected separators/classifiers on simulated data sets from field of particle physics

Marcel Jiřina and František Hák

Technical Report No. V-1089

November 2010

Abstract

Results given in this report show that the minimal classification error and the “left part” of the ROC curve are very different things. The “left part” of the ROC curve corresponds to the highest rejection factor that is needed in data processing for particle physics. It is shown that one can suppress all background events for limited, even though large, number of background events and that the threshold for cutting off background events depends on the number of events considered for *test*. The most important message is that testing of separation techniques should be done with realistic numbers of events of both classes in the *testing* set.

Keywords:

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

Testing selected separators/classifiers on simulated data sets from field of particle physics

Marcel Jiřina and František Hák

Contents	
Introduction	3
Data sets description	3
“Juránek 16“ and “Juránek New 16“	3
Original data set	3
New data set	3
“Elsbieta 7“ and “Elsbieta 25“	4
25 variables set	4
Seven variables data set	4
“Magic“	4
“Řezníček“	5
Classifiers/separators used	6
Results	7
Summary of results	7
“Juránek 16“	8
“Juránek New 16“	9
“Elsbieta 25“	10
“Elsbieta 7“	11
“Magic“	12
“Řezníček“	13
Discussion	14
Acknowledgement	14
References	14

Introduction

In year 2010 we tested and compared several very different classifiers using data sets at hand and mostly used in previous analyses. Results show that for some data a better separation of signal and background can be reached using a different tool. Also, very different things are the minimal classification error and the “left part” of the ROC curve. The minimal classification error corresponds to the overall course of a ROC curve. The “left part” of the ROC curve corresponds to the highest *rejection factor* that is needed in data processing for particle physics. It is shown here that it is possible to suppress all background events for limited, even though large, number of background events. It is also shown here that the threshold for cutting background events¹ depends on the number of events considered for test and not only on the number of events contained in the learning set.

Data sets description

“Juránek 16” and “Juránek New 16”

Mr. Juránek generated this data in 2009. The signal and background are so called exclusive diffraction processes where protons will remain as a particle even after collision. They interact so that they interchange some colorless object (two gluons). It means that all energy that is interchanged (i.e. energy of that gluons) is transformed into some so-called central object as e.g. Higgs boson, a pair quark-antiquark, a pair of gluons or some another pair of particles. In this case the central object considered is the Higgs boson that decays next into a pair of quarks $b - b\text{-bar}$ and the background process where the central object is a pair of quarks $b - b\text{-bar}$. The importance of this process lays in the fact that standard (no diffraction) production of the Higgs boson decaying into $b - b\text{-bar}$ pair is impossible to measure in LHC due to extremely large QCD $b - b\text{-bar}$ background.

Original data set

It consists of 16 variables as follows:

Prot1E; Prot1_{px}; Prot1_{py}; Prot1_{pz};
Prot2E; Prot2_{px}; Prot2_{py}; Prot2_{pz};
Jet1E; Jet1_{px}; Jet1_{py}; Jet1_{pz};
Jet2E; Jet2_{px}; Jet2_{py}; Jet2_{pz}.

i.e. there are four „four-jets“ for two protons and two jets measured.

File names are lrn0.dta and tst0.dta for learning and testing set, respectively.

New data set

The “New” data set has the same 16 variables as the original data set but a different system of cuts has been previously applied. Thus the signal and background events in this data are more difficult to recognize.

File names are lrn.dta and tst.dta for learning and testing set, respectively.

¹ This is analogous but has nothing to do with “cuts” approach widely used in particle physics. Most of separators/classifiers give a single scalar output value as a response to multivariate input data sample (event). A proper threshold then differentiates between classes, i.e. what is above this threshold is recognized as a signal, what lies below is a background. (Or vice versa depending on construction of the classifier.)

“Elsbieta 7“ and “Elsbieta 25“

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 $H^\pm \rightarrow \tau\nu$ 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).

25 variables set

In this set a more variables i.e. a more detailed description of decay processes are used. The 25 variables consist of six „four-jets“ and lepton. The learning and testing data sets L-data_05_05_30_09_11_21.dat and T-data_05_05_30_09_11_21.dat

Seven variables data set

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 describe them in detail here as it may be found in [4]. 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.

“Magic“

For description of data we cite [5] here verbatim as follows: “Ground-based atmospheric Cherenkov telescopes using the imaging technique are a comparatively recent addition to the panoply of instruments used by astrophysicists. The first results were demonstrated in 1989. They observe high-energy gamma rays, taking advantage of the radiation emitted by charged particles as they are produced abundantly inside the electromagnetic showers initiated by the gammas, and developing in the atmosphere. This Cherenkov radiation (of visible to UV wavelengths) leaks through the atmosphere and gets recorded in the detector, allowing reconstruction of shower parameters.

For our case study, we used data sets generated by a Monte Carlo program, Corsika, described in ref. The program was run with parameters allowing to observe events with energies down to well below 50 GeV.

Subsequently, the analysis is simplified, with hopefully little or no loss of information, by converting the pixel image of a shower into few image parameters as described earlier. These parameters constitute the only image characteristics to be used.

The data consist of two classes: gammas (signal) and hadrons (background). Events were generated at shower energies from 10 GeV up to about 30 TeV, and for zenith angles from zero to 20 degrees. The samples used by all methods are identical, and consist of 12332 gamma events and 6688 hadron events. Each event is characterized by the following ten parameters:

1 length : major half axis of ellipse [mm]

2 width: minor half axis of ellipse [mm]
 3 size: 10-log of sum of content of all pixels [photon count]
 4 conc2 : ratio of sum of two highest pixels over size [ratio]
 5 conc1 : ratio of brightest pixel over size [ratio]
 6 pdist: distance from brightest pixel to center, along major axis [mm]
 7 m3long: 3rd root of third moment along major axis [mm]
 8 m3trans: 3rd root of third moment along minor axis [mm]
 9 alpha: angle of major axis with vector to origin [deg]
 10 dist: distance from origin to center of ellipse [mm]
 All multivariate methods studied here use identical disjoint training (learning) and control (test) samples. The data sets used are magic.dat and magic.dat for learning and testing set, respectively.

“Řezníček”

This data is quite large consisting of 109 289 events. There are four original data sets signal_small.txt, background_small.txt, and signal_big.txt, background_big.txt. There are 9790 signal events and 99499 background events.

The first pair consists of 54 variables that were mostly used in other analyses. The second pair of data has 68 variables more, i.e. total 122 variables. Additional variables can improve classification, but, on the other hand, are highly correlated with the basic set of 54 variables. Twelve variables of that set of 54 variables are integers, but it makes no problem.

In detailed analysis of individual variables we found that four of them bring no new information that resulted in the use of 118 variables only.

Tests were performed with “big” data, i.e. data with resulting 118 variables. Thus we tested ability of classifiers to process data with more than 100 variables.

The learning data file Lear1kB.txt has 1000 events, 500 of background, and 500 of signal. The testing data file Test1kB.txt consists of all remaining events, and short version Test1kB10k.txt has 10000 events (5000 signal events, 5000 background events). In Results section The short version is denoted “Rezn10k”, long version simply “Reznicek”. In both cases the learning set is the same, the Lear1kB.txt. In some tests the testing data set was limited to 100000 events by deleting some background events.

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 18 different classifiers/separators and their variants as follows

Method	Description
10-bins Bayes – naïve	Classical Bayes naïve classifier that uses ten-bins histograms of individual variables and not any approximation by a distribution density function.
Chi2 combined with IINC	A special method that generates two-dimensional maps of events for both classes [6]. Classes are then separated by the use of simple IINC method.
1-NN method L1	Nearest neighbor method [3] with L1 (Manhattan) metrics
1-NN method L2	The same with L2 (Euclidean) metrics
5-NN method L1	Five nearest neighbors method [3] with L1 (Manhattan) metrics
5-NN method L2	The same with L2 (Euclidean) metrics
0-NN method (sqrt(n)) L1	Nearest neighbors method with neighborhood size given by square root of the number of samples of the learning set with L1 (Manhattan) metrics
0-NN method (sqrt(n)) L2	The same with L2 (Euclidean) metrics
IINC (1/I) method L1	Inverse Indexes of Neighbors Classifier [7], [8]. Relatively simple method derived on the bases of estimating multifractal dimensions (Hurst exponents) and Zipfian distribution. Here with L1 (Manhattan) metrics
IINC (1/I) method L2	The same with L2 (Euclidean) metrics
Qcregre standard L1	Method that uses a constant when a multifractal dimension (Hurst exponent, distribution mapping exponent) is computed for a given event (sample) [9]. Here with L1 (Manhattan) metrics
Qcregre standard L2	The same with L2 (Euclidean) metrics
DME-local standard L1	Distribution mapping exponent method [10] that polynomially transforms a true probability distribution to be locally uniform and then estimates a class to which an event belongs. With L1 (Manhattan) metrics
DME-local standard L2	The same with L2 (Euclidean) metrics
CD-global standard L1	Method that polynomially transforms a true probability distribution to be uniform-like. For it uses (global) correlation dimension as an exponent and then estimates a class to which an event belongs [11]. With L1 (Manhattan) metrics
CD-global standard L2	The same with L2 (Euclidean) metrics
Random Forest (RandFor)	The well-known method by Leo Breiman and Adele Cutler [1], [2]

Results

Summary of results

The summary is given in the following Table. Note that Table below gives the overall minimal classification error in individual entries. The classification error is maximized without respect to the value of the rejection factor really needed in evaluating events. At the same time, it sometimes happens that a classifier becomes stupid interchanging classes and thus giving classification error close to 50 % that would correspond to purely random decision.

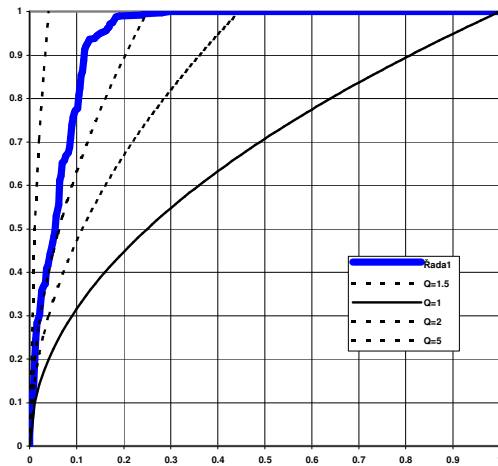
Method	Metrics	D a t a s e t							Method
		Jur16	Jur New16	Elsb25	Elsb7	Magic	Rezn10k	Reznice k	Mean
10-bins Bayes - naive	NA	25.77%	49.70%	40.28%	30.84%	16.68%	1.29%	1.76%	23.84%
Chi2 combined with IINC	L2	24.74%	43.50%	43.07%	27.94%	38.34%	2.34%	0.69%	25.80%
1-NN method	L1	36.48%	49.70%	43.33%	28.37%	32.59%	4.81%	9.16%	29.28%
1-NN method	L2	40.82%	49.60%	44.85%	28.28%	33.69%	5.49%	8.95%	30.40%
5-NN method	L1	34.95%	49.10%	47.65%	24.44%	26.08%	5.29%	8.80%	28.28%
5-NN method	L2	36.73%	48.90%	45.11%	23.61%	26.47%	5.80%	8.93%	28.21%
0-NN method (sqrt(n))	L1	32.91%	48.40%	41.68%	22.21%	22.70%	5.15%	9.78%	26.21%
0-NN method (sqrt(n))	L2	35.20%	49.60%	42.57%	21.72%	22.80%	5.53%	12.08%	27.38%
IINC (1/i) method	L1	33.80%	49.60%	41.42%	22.70%	28.01%	5.76%	8.45%	26.97%
IINC (1/i) method	L2	35.46%	49.20%	37.99%	22.12%	28.79%	6.88%	9.50%	26.93%
QCrege standard	L1	33.67%	47.40%	43.96%	21.51%	17.06%	5.59%	9.89%	25.58%
QCrege standard	L2	36.10%	49.40%	45.24%	21.75%	20.20%	4.51%	6.42%	26.23%
DME-local standard	L1	34.06%	49.10%	40.79%	23.12%	24.65%	4.96%	8.55%	26.46%
DME-local standard	L2	36.48%	50.00%	39.14%	23.12%	25.07%	5.86%	10.15%	27.12%
CD-global standard	L1	37.76%	49.80%	39.77%	35.66%	28.79%	NA	NA	38.43%
CD-global standard	L2	38.65%	48.90%	39.52%	36.09%	29.75%	NA	NA	39.02%
RandomForest	NA	9.95%	38.40%	37.23%	21.54%	16.01%	0.85%	1.02%	17.86%
Data set mean error (measures difficulty):	NA	33.18%	48.90%	43.88%	26.76%	25.75%	4.67%	7.13%	28.62%

In this table the last column gives the “Method Mean”. It is the mean value of the classification error over all seven data sets (problems) considered. The smaller the Method Mean the better method from general point of view. The last row in the table shows the Data set mean error. This value is the mean classification error over all classification methods including their variants for a particular data set. Comparing values on the last row of the table we see that data set Rezn10k appears most easily to separate. On the contrary the data set Jur New 16 appears as the most difficult to separate.

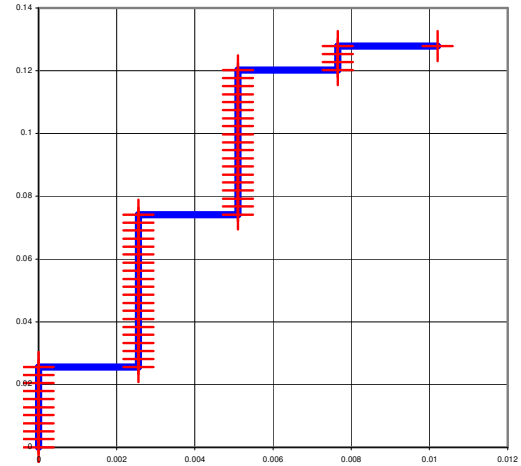
All distance-based methods are considered in two variants according to metrics used. We tested them with L1 (Manhattan) metrics and with standard L2 (Euclidean) metrics. Small differences show a slight advantage of L1 metrics over L2 metrics in most cases.

“Juránek 16”

Original data set.



ROC curve for data “Juránek 16”, smoothed.



ROC curve for data “Juránek 16”, not smoothed, the left end detail. Red crosses indicate individual events.

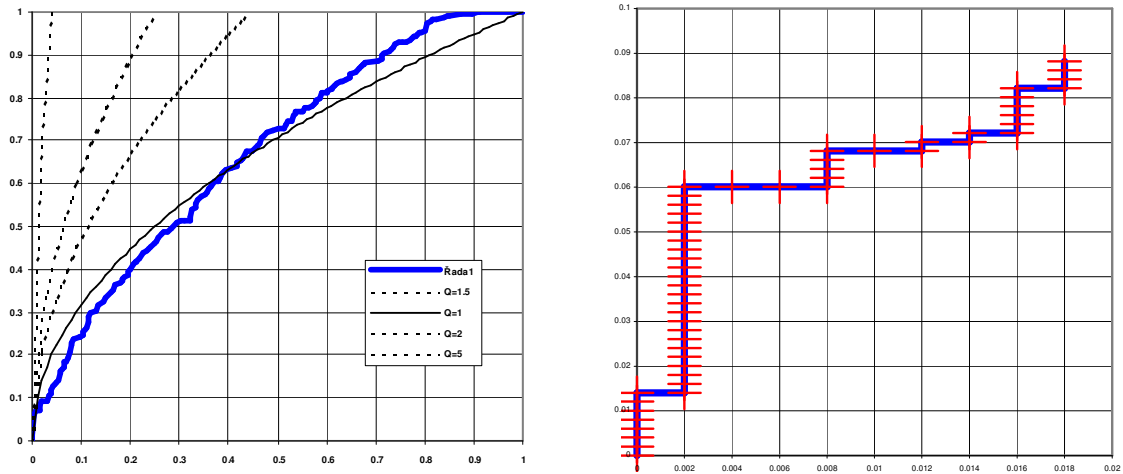
In a more detailed analysis one can find results according to table:

Background events after separator	0	1	2
Rejection factor	∞	392	196
Background error	0	0.002551	0.0051
Signal efficiency	0.025	0.073	0.104
Threshold	0.94	0.921	0.903

This means e.g. that for threshold 0.94 all background events can be rejected and, at the same time we get 2.5 % of original signal events. When we wish to get at least 10 % of signal events then we must accept 0.5 % of background events going through separator (now with threshold 0.903).

“Juránek New 16”

The data set but with heavier cuts applied previously.



ROC curve for data “Juránek New 16”, not smoothed. ROC curve for data “Juránek New 16”, not smoothed, the left end detail. Red crosses indicate individual events.

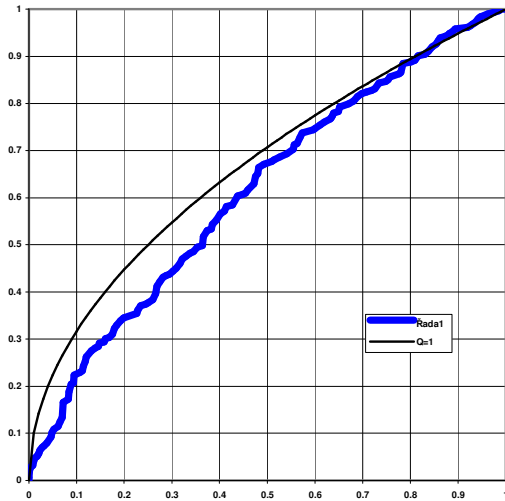
In a more detailed analysis one can find results according to table:

Background events after separator	0	1	4
Rejection factor	∞	500	125
Background error	0	0.0020	0.008
Signal efficiency	0.012	0.58	0.066
Threshold	0.801	0.743	0.732

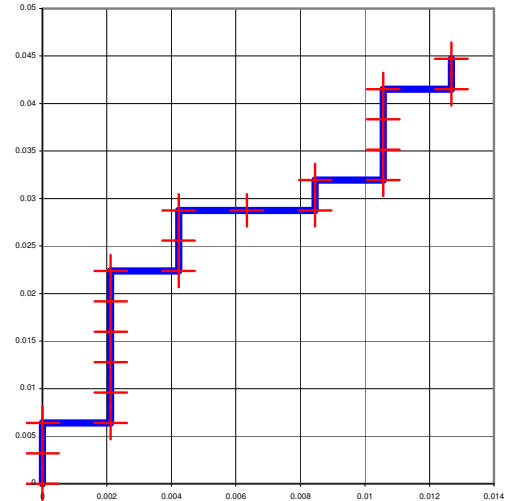
This means e.g. that for threshold 0.801 all background events can be rejected and, at the same time we get 1.2 % of original signal events. When we wish to get at least 6.6 % of signal events then we must accept 0.08 % of background events going through separator (now with threshold 0.732).

“Elsbieta 25”

25 variables set. It is extremely difficult to separate signal from background in this data as seen in the following Figure.



ROC curve for data “Elsbieta 25”, smoothed.



ROC curve for data “Elsbieta 25”, not smoothed, the left end detail. Red crosses indicate individual events.

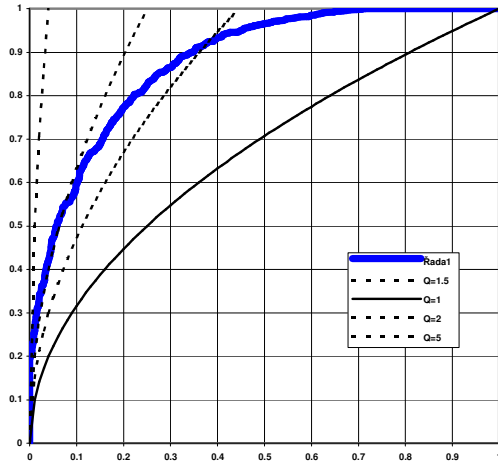
In a more detailed analysis one can find results according to table:

Background events after separator	0	1	2
Rejection factor	∞	473	236.5
Background error	0	0.002114	0.004228
Signal efficiency	0.00319	0.0191	0.025556
Threshold	0.728	0.700	0.660

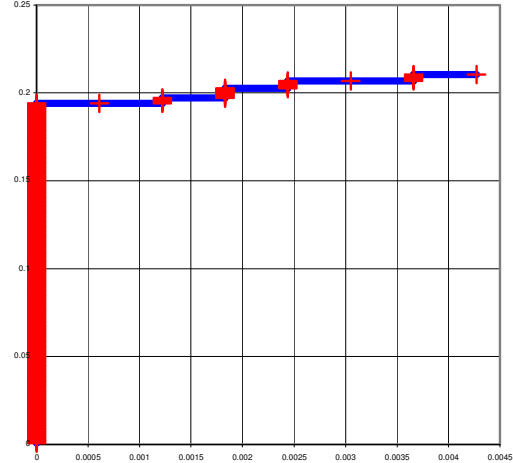
This means e.g. that for threshold 0.728 all background events can be rejected but at the same time we get 0.319 %, i.e. approximately one of 300 of signal events only. When we wish to get, say, at least 2.5 % of signal events then we must accept 0.4228 % of background events going through separator (with threshold 0.660).

“Elsbieta 7”

Seven variables data set. This data is interesting by the fact, that the best separator is the “QCrege” method that uses a constant when a multifractal dimension is computed for a given event with L1 (Manhattan) metrics.



ROC curve for data “Elsbieta 7”, smoothed.



ROC curve for data “Elsbieta 7”, not smoothed, the left end detail. Red crosses indicate individual events.

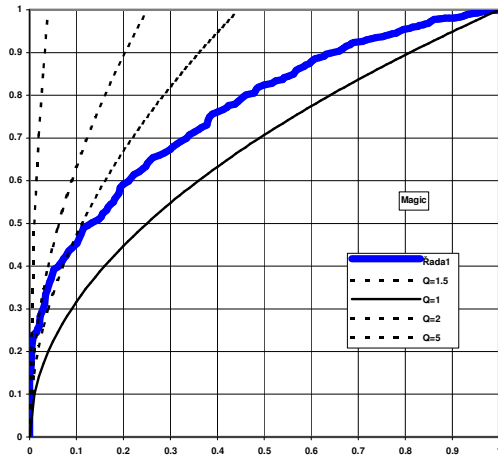
In a more detailed analysis one can find results according to table:

Background events after separator	0	2	3
Rejection factor	∞	819.5	546
Background error	0	0.00122	0.00183
Signal efficiency	0.193	0.196	0.201
Threshold	0.954	0.902	0.869

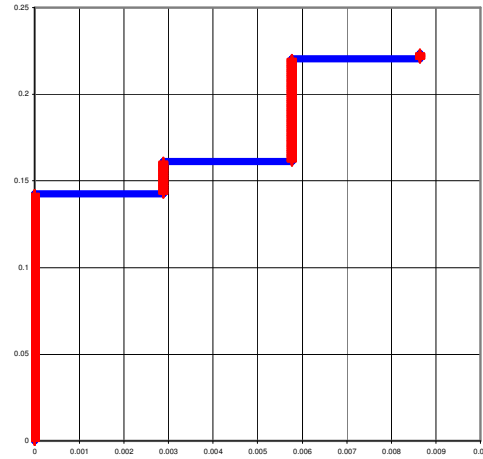
This means e.g. that for threshold 0.954 all background events can be rejected and, at the same time we get 19.3 % of original signal events. The ROC curve is rather flat in region considered, see Figure above at the right hand side. Then lowering the threshold helps a little to get more signals and causes a rejection factor less than 1000.

“Magic”

The data set is available at the UCI Machine Learning Repository.



ROC curve for data “Magic”, smoothed.



ROC curve for data “Magic”, not smoothed, the left end detail. Small red crosses indicate individual events.

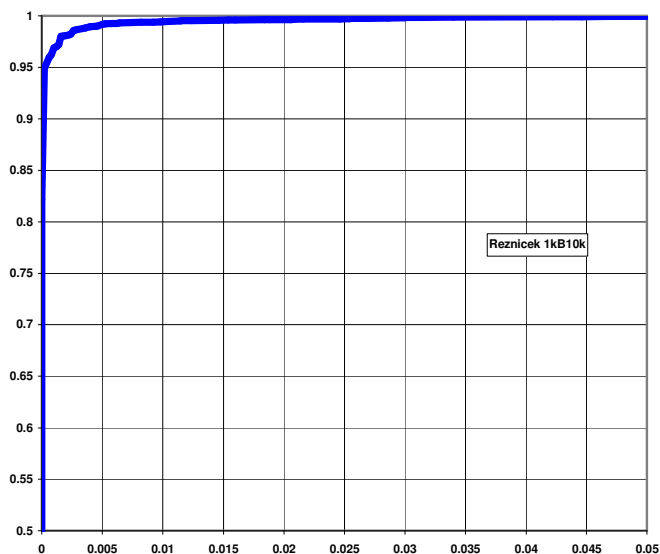
In a more detailed analysis one can find results according to table:

Background events after separator	0	1	2
Rejection factor	∞	347	173.5
Background error	0	0.002882	0.005764
Signal efficiency	0.142	0.161	0.22
Threshold	0.805	0.787	0.737

This means e.g. that for threshold 0.805 all background events can be rejected and, at the same time we get 14.2 % of original signal events. When we wish to get 22 % of signal events then we must accept 0.576 % of background events going through separator (now with threshold 0.737).

“Řezníček”

This data is interesting by nearly “ideal” ROC curve. Signal and background events are relatively easy to separate.



The upper left (!) corner of the ROC curve for data “Řeznicek 1kB10k”, smoothed.

In a more detailed analysis one can find results according to table:

10 000 testing events data set				
Background events after separator	0	2	3	4
Rejection factor	∞	2500	1666.667	1250
Background error	0	0.0004	0.0006	0.0008
Signal efficiency	0.9494	0.9604	0.9624	0.9694
Threshold	0.831	0.795	0.789	0.767

This means e.g. that for threshold 0.831 all background events can be rejected and, at the same time we get 94.9 % of original signal events.

For all data (109269 events) results are summarized in the following table.

All 109 269 testing events data set				
Background events after separator	0	1	2	3
Rejection factor	∞	99499	49749.5	33166.3
Background error	0	1.01E-05	2.01E-05	3.02E-05
Signal efficiency	0.58223	0.62809	0.66834	0.83207
Threshold	0.995	0.993	0.991	0.963

This means e.g. that one can get from total 109289 events (9790 sign., 99499 backg.) 4978 signals and no background, i.e. all background events can be rejected and 50.8 % of signal events remain. At the cost of nonzero background error equal to 0.00003 (0.003 %) one can obtain 83.2 % signal efficiency.

It is interesting that classifier better in the sense of giving a smaller overall classification error gives worse results in the same region as follows.

100 000 testing events data set				
Background events after separator	0	1	2	3
Rejection factor	∞	90210	45105	30070
Background error	0	1.1E-05	2.2E-05	3.3E-05
Signal efficiency	0.48764	0.49612	0.52431	0.55312
Threshold	0.62395	0.62049	0.61077	0.40032

As seen, results are roughly by ten per cent worse than in the previous case.

Discussion

Even generally the best classifier can be „beaten“ by another method in case of a particular data set. In such a case one can find that rather simple algorithm can outperform a highly sophisticated one. From it follows that one has to have a set of classifiers. For particular data one should optimize results first by selection a proper classifier. Then eventually optimize its parameters, as usually default parameters are rather close to those needed for the best result.

Comparing tables in Chap. Dealing with data “Řezniček” for small (10 000 events) and large (108289 or 100 000 events) testing data sets and the same separator/classifier we see that the threshold is shifted to larger value for larger data set. It means more severe cutting of background events (that is ok) but at the same time also more severe cutting of signal events. Thus keeping background events eliminated results in lower acceptance (efficiency) of signal events. Fortunately it is not linear. Here background eliminated is ten times larger whereas signal events acceptance is reduced approx. to half only.

The most important message is that testing of separation techniques should be done with realistic numbers of events of both classes in the testing set.

In an opposite case one cannot extrapolate results until some asymptotic to realistic case is found. Not to be so strict, the best separation tool for a particular case remains the best one for different data set size with high probability.

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] Breiman,L.: Random Forests, Machine Learning Vol. 45, No. 1, pp. 5-32 (2001)
- [2] Jirina,M., Jirina,M.,jr.: Testing Random Forest for Unix and Windows. Technical Report No. V-1075, Institue of Informatics AS CR, Prague (2010)

- [3] T. M. Cover and P. E. Hart. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, vol. IT-13, No. 1, pp. 21-27, (1967)
- [4] Hakl,F., Jirina,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)
- [5] R. K. Bock et al.: Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. Nuclear Instruments and Methods in Physics Research vol. A 516, pp.511-528, (2004)
- [6] F. Bečvář (MFF UK Prague): Personal communication.
- [7] M Jirina, M. Jirina 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.
- [8] M Jirina, M. Jirina 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.
- [9] M Jirina, M. Jirina, jr.: Probability Density Estimation by Decomposition of Correlation Integral. Proc of the 2008 International Conference on Artificial Intelligence and Pattern Recognition (AIPR-08) Orlando, Florida, USA, July 7-10, 2008, B. Prasad, P.Sinha (Eds.), pp. 113-119 (paper No. AIPR165).
- [10] Jiřina, Marcel. Distribution Mapping Exponent for Multivariate Data Classification. In Computer Science and Engineering 5. Orlando, US, 2004. pp. 103-108. ISBN 980-6560-13-2. [SCI 2004. World Multi-Conference on Systemics, Cybernetics and Informatics /8./, Orlando, 18.07.2004-21.07.2004, US].
- [11] Jiřina, Marcel ; Jiřina jr., M. Simple and Effective Probability Density Estimation and Classification. In SICE-ICCAS 2006. Seoul: National University Press, 2006. s. 4479-4480. ISBN 89-950038-5-5. [SICE-ICASE International Joint Conference 2006, Busan, 18.10.2006-21.10.2006, Korea]