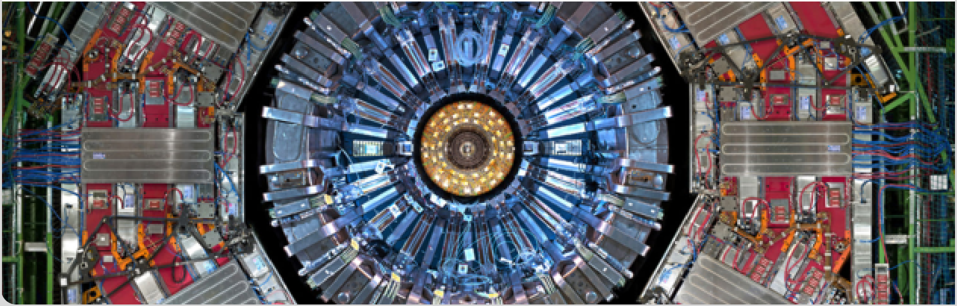
Comparison of multiclassification frameworks in the context of the $t\bar{t}H(b\bar{b})$ analysis

Patrick Golz | 14.11.2018

INSTITUT FÜR EXPERIMENTELLE TEILCHENPHYSIK (ETP)

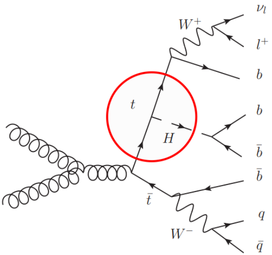


KIT – The Research University in the Helmholtz Association

www.kit.edu

Motivation:

- Check Standard Model nature of Higgs boson
- Measurement of Yukawa coupling
  $\rightarrow$ Possible probe for new physics
- Coupling strength proportional to fermion mass → Top-Higgs coupling
- Direct access to coupling → $t\bar{t}H$
- $H \rightarrow b\bar{b}$ has largest branching ratio (58 %) → $t\bar{t}H(b\bar{b})$
- Semileptonic channel: balance between background rejection and statistics

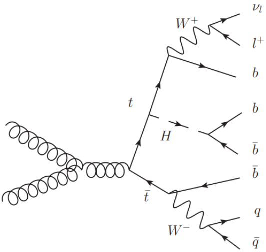


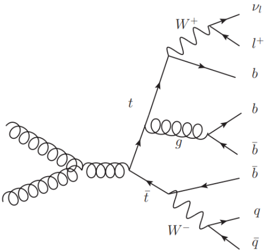
Difficulties

- Very similar final states of $t\bar{t}H$ and $t\bar{t}$ + (b) jets background processes
- $t\bar{t}$ + (b) jets background exceeds $t\bar{t}H$-signal significantly ($\sigma_{t\bar{t}b\bar{b}} \approx 5\,\text{pb}$ vs. $\sigma_{t\bar{t}H} = 0.51\,\text{pb}$)

→ Use of multivariate analysis methods necessary

→ Multiclassifiers especially promising





Output classes

Multiclassification has been done in a previous work by L. Hilser using following classes:

- $t\bar{t}H(b\bar{b})$: signal
- events in which a $t\bar{t}$ pair and a $b\bar{b}$ pair are created, are separated into three classes:
    - $t\bar{t}b\bar{b}$: events in which both bottom quarks are registered as separated jets
    - $t\bar{t}2b$: events in which both b jets strongly overlap and can not be separated
    - $t\bar{t}b$: events in which only one of the two bottom quarks is detected
- $t\bar{t}c\bar{c}$: events with a $t\bar{t}$ pair in connection with at least one additional jet containing at least one charmlike hadron
- $t\bar{t}$+LF: events in which a $t\bar{t}$ pair in connection with lighter quark jets is created

Dataset

- Monte-Carlo events created for $t\bar{t}H$ analysis in 2016 with Powheg + Pythia 8
  - $t\bar{t}H(b\bar{b})$
  - $t\bar{t}$
- Center of mass energy: 13 TeV
- Preselection cuts:
  - 6 or more jets with $P_t \geq 30\,\text{GeV}$
  - 2 or more b-tagged Jets
- 800 000 events after preselection

Problem: TensorFlow can not work with ROOT files directly
$\rightarrow$ Are there more suitable frameworks available?

Frameworks

- TensorFlow: Framework for artificial neural networks
- Keras: easy to use high level API built on TensorFlow
- NNFlow:
  - Basic TensorFlow script for use in our analysis
  - Developed by M. Welsch, M. Lang and L. Hilser at ETP
- TMVA:
  - Multivariate analysis toolkit for ROOT
  - Provides Keras interface and internal DNN implementation

Frameworks

NNFlow has been used for multiclassification previously by L. Hilser
→ Comparison of NNFlow and TMVA regarding performance and usability

Used TMVA methods:

- TMVA-DNN:
  - DNN directly implemented in TMVA
  - Optimized for use with ROOT files
- TMVA-Keras: Keras interface for TMVA
- TMVA-BDTG: gradient boosted decision tree

Configuation of NNFlow-DNN

The DNNs in TMVA have been modeled as close as possible after the NNFlow DNN
→ comparability

hidden layer layout	100, 100
output nodes	6
input features	10 high level variables
activation function	elu
dropout probability	0.3
L2 regularisation	10^{-12}
early stopping interval	15
optimizer	Adam
batch size	500

Calibration options

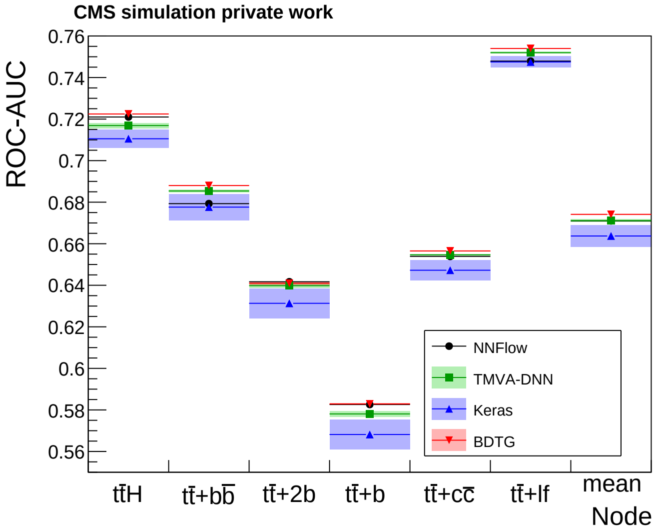
- Versions: ROOT 6.12 / TMVA 4.3 (12.12.2017)
- TMVA-DNN has fewer options for activation functions and optimizers
- TMVA has very restricted options to analyze the training process
- NNFlow can be easily edited
  → Most of TFs options can be accessed

Training time

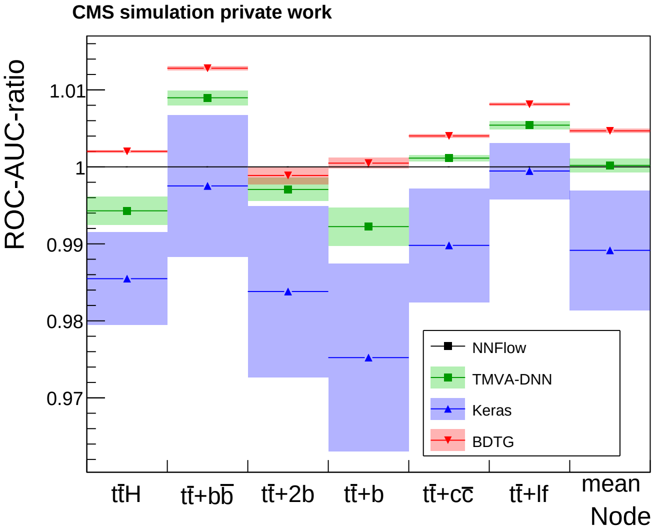
	overall time in s	time per epoch in s	evaluation time in μs
NNFlow	731 ± 138	2,344 ± 0,013	-
TMVA-DNN	165 ± 1	0,244 ± 0,002	18 ± 1
TMVA-Keras	1016 ± 369	2,370 ± 0,043	467 ± 6
TMVA-BDTG	2165 ± 22	-	220 ± 8

- to find the right hyperparameters, training has to be repeated many times
  → faster training times is a big advantage

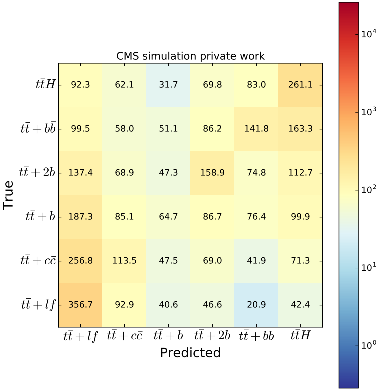
Absolute ROC values



Relative deviation of ROC values

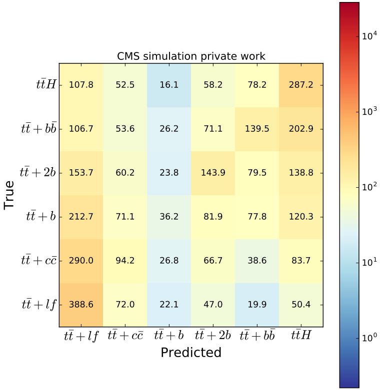


Confusion matrices



	$t\bar{t}+lf$	$t\bar{t}+c\bar{c}$	$t\bar{t}+b$	$t\bar{t}+2b$	$t\bar{t}+b\bar{b}$	$t\bar{t}H$
$t\bar{t}H$	92.3	62.1	31.7	69.8	83.0	261.1
$t\bar{t}+b\bar{b}$	99.5	58.0	51.1	86.2	141.8	163.3
$t\bar{t}+2b$	137.4	68.9	47.3	158.9	74.8	112.7
$t\bar{t}+b$	187.3	85.1	64.7	86.7	76.4	99.9
$t\bar{t}+c\bar{c}$	256.8	113.5	47.5	69.0	41.9	71.3
$t\bar{t}+lf$	356.7	92.9	40.6	46.6	20.9	42.4

(a) NNFlow



	$t\bar{t}+lf$	$t\bar{t}+c\bar{c}$	$t\bar{t}+b$	$t\bar{t}+2b$	$t\bar{t}+b\bar{b}$	$t\bar{t}H$
$t\bar{t}H$	107.8	52.5	16.1	58.2	78.2	287.2
$t\bar{t}+b\bar{b}$	106.7	53.6	26.2	71.1	139.5	202.9
$t\bar{t}+2b$	153.7	60.2	23.8	143.9	79.5	138.8
$t\bar{t}+b$	212.7	71.1	36.2	81.9	77.8	120.3
$t\bar{t}+c\bar{c}$	290.0	94.2	26.8	66.7	38.6	83.7
$t\bar{t}+lf$	388.6	72.0	22.1	47.0	19.9	50.4

(b) TMVA-DNN

Performance results

- Considering the random fluctuations, all classifiers perform equally well
- Largest differences seen in confusion matrices of the TMVA-DNN → probably due to different optimizer and activation functions
- Very similar performance of the BDT compared to the DNNs

Conclusion

- TMVA methods achieve similar results to NNFlow and can be used as a viable alternative
- Advantages:
    - easy to use
    - can use ROOT files directly
    - can be used with C++
      → easy to integrate into the workflow
    - considerably shorter training times for TMVA-DNN
    - easy to compare different classifiers → perfect for quickly testing new ideas
- Disadvantages:
    - configuration options are limited

Backup

Configuation of NNFlow-DNN

hidden layers	100, 100
activation function	elu
dropout probability	0.3
L2 regularisation	10^{-12}
early stopping interval	15
optimizer	Adam
β_1	0.9
β_2	0.999
ϵ	10^{-8}
learning rate	$3 \cdot 10^{-5}$
learning rate decay	deactivated
batch size	500

Configuration of TMVA-Keras

hidden layers	100, 100
activation function	elu
dropout probability	0.3
L2 regularisation	10^{-12}
early stopping interval	10
optimizer	Adam
β_1	0.9
β_2	0.999
ϵ	10^{-8}
learning rate	$3 \cdot 10^{-6}$
learning rate decay	deactivated
batch size	500

Configuration of TMVA-DNN

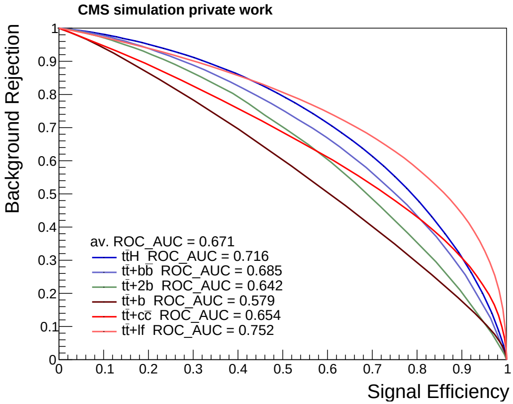
hidden layers	100, 100
activation functions	relu
dropout probability	0.7
early stopping interval	10
learning rate	10^{-2}
momentum	deactivated
L2 regularisation	deactivated
batch size	500

Configuration of TMVA-BDTG

number of trees	1000
boosting	Gradient Boosting
shrinkage	0.1
minimal node size	1%
bagged boosting	active
bagged sample fraction	0.5
number of cuts	20
max depth	2

ROC curves for multiclassification are calculated as follows:

$$(x, y) = \left(\frac{1}{N}\sum s, \frac{1}{N_b}\sum_i \frac{1}{N_i}\sum b_i\right) \tag{1}$$

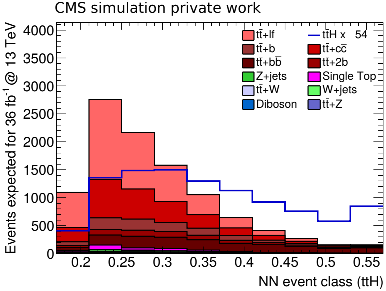


Input features

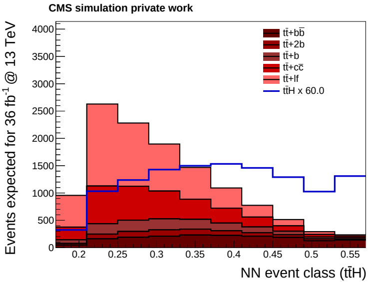
- ΔR between two b-tagged jets mit b-Tag, mean over all possible combinations
- mass of vector sum of the lepton and the b-tagged jets with smallest ΔR to the lepton
- CSV value of the jet with the second highest csv value
- specifies if an event contains more likely 4 or 2 b-jets
- mean CSV value of all b-tagged jets
- variance of the CSV values
- $\Delta\eta$ between the two b-tagged jets with the smallest ΔR
- ΔR between the lepton and the jet with the smallest ΔR to the lepton
- ΔR between the two b-tagged jets with the smallest ΔR
- mass of vector sum of two b-tagged jets, mean over all possible combinations

$\Delta\Phi$: difference between azimuth angles; $\Delta\eta$ difference between pseudorapidities; $\Delta R = \sqrt{\Delta\Phi^2 + \Delta\eta^2}$

Discrimination plots



(a) NNFlow



(b) TMVA-DNN