Modern Methods in Data Analysis

Multivariate Data Analysis: Higgs Challenge

Lukas Fritz	1686473
Fabian Leven	1638446
Ali Deniz Özdemir	1724032
Lena Salfenmoser	1723697
Johannes Heizmann	1725035

4. Februar 2015

Inhaltsverzeichnis

1 Introduction

[?]

2 Auswahl einer geeigneten Untermenge an Parametern

2.1 Vorgehen

Um eine geeignete Untermenge an Parametern zu finden, haben wir die Relevanz der Parameter evaluiert. Dafür sind wir wie folgt vorgegangen:

Die TMVA-Methoden geben selbst eine Ranglistein Bezug auf die Relevalnz der Parameter aus. Um diese zu evaluieren, liesen wir die Trainingsalgorithmen dreisig mal durchlaufen. Jedes Mal mit einer Variable weniger und die Reihenfolge des Entfernens entsprach der Rangliste. Jedes Mal wurde der AMS-Wert protokoliert. Diese Methode werteten wir für das Liklihood-, das Fisher- und das BDT-Verfahren aus.

Für das MLP-Verfahren gingen wir wie folgt vor: Wir trainierten das neuronale Netz dreisig mal und liesen jedes Mal einen anderen Parameter weg. Desto höher der erreichte AMS-Wert war, desto irrelevanter bzw. sogar nachteiliger ist der weggelassene Parameter. Gemäß des AMS-Werts lässt sich ebenfalls eine Rangliste erstellen. Danach wurde so verfahren, wie im vorangehenden Abschnitt beschrieben: Das Training wurde mit 30, dann mit 29, dann mit 29, ... Parametern durchgeführt, wobei die Reihenfolge des Weglassens der erstellten Rangliste entsprach.

Das beste vorgehen war allerdings Folgendes: Wir trainierten die Methoden und liesen dabei jedes Mal einen anderen Paramerter weg. Der Parameter dessen Entfernen den höchsten AMS-Wert erreichte wurde aus der Parametermenge entfernt. Diesen Vorgehen wurde wiederholt bis nur noch ein Parameter übrig war. Dafür sind 465 Trainingszyklen notwendig. Aufgrund der langen Trainingszeit haben wir dieses Vorgehen nicht bei der MLP-Methode verwendet.

2.2 Auswertung

Die Ergebnisse sind in Abbildung 1 gezeigt. Welche Parameteranzahl wir für jede Methode verwendet haben ist durch einen senkrechten Strich gekennzeichent und liegt jeweils beim maximalen AMS-Wert. Die Rangliste der Parameter und welche Parameter wir letztendlich benutzt haben, steht in Tabelle 1.

Darüber hinaus sieht man, dass bei der Likelihood- und bei der Fisher-Methode unsere Rangliste der Parameter ein besseres Ergebnis liefert als die interne Rangliste der TMVA-Methoden.

Ausschlaggebend bezüglich des AMS-Wertes ist die Verkleinerung der Parametermenge lediglich bei der Likelihood-Methode. Allerdings bringt eine reduzierte Parametermenge auch eine reduzierte Laufzeit des Algorithmus mit sich, was insbesondere für die MLP-Methode nützlich ist, da sie eine lange Trainigszeit erfordert.

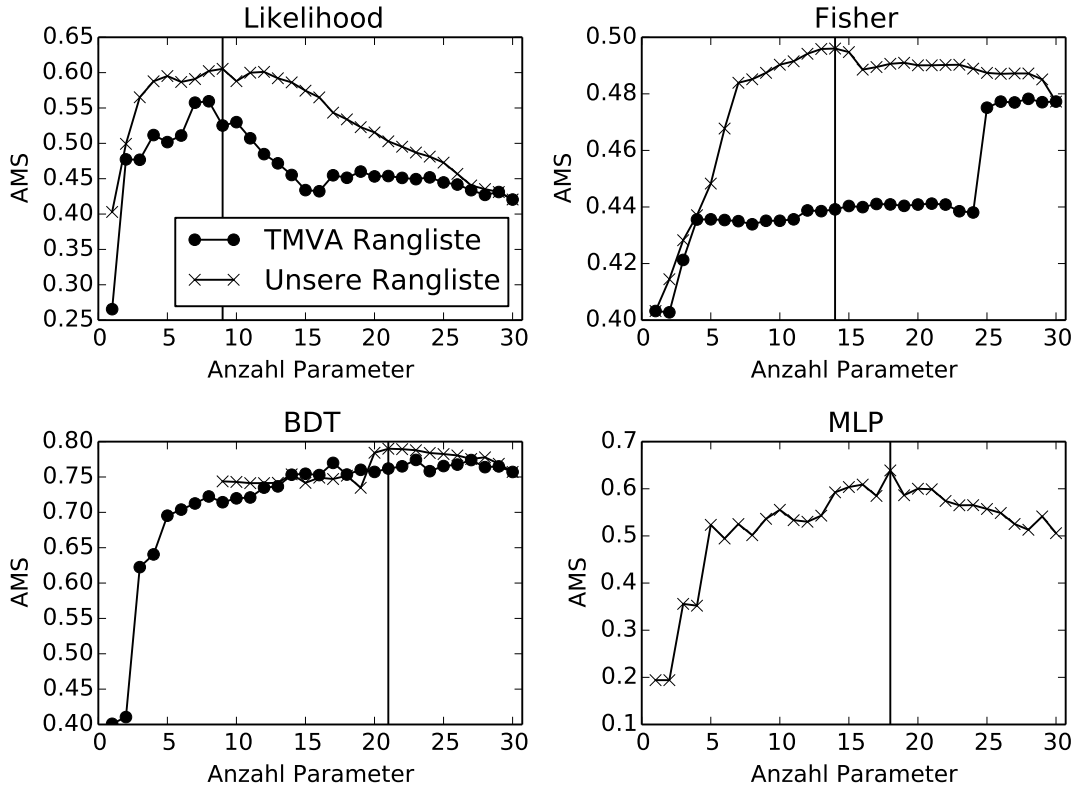

Abbildung 1: AMS über der Anzahl der Parameter.

Tabelle 1: Bewertung der Parameter. Für die einzelnen Methoden nimmt die Gewichtung der Parameter von oben nach unten ab. Für jede Methode ist durch „——“ gekennzeichnet, welche Parameter nicht mehr verwendet werden. Bei der BDT-Methode wurde unter den besten acht Parametern kein Ranking mehr vorgenommen.

Likelihood	Fisher	BDT	MLP
d_mass_transverse_met_lep	d_mass_transverse_met_lep	-	p_jet_num
d_mass_vis	d_pt_ratio_lep_tau	-	d_deltaeta_jet_jet
p_tau_pt	p_lep_pt	-	d_mass_MMC
d_deltar_tau_lep	d_mass_vis	-	p_jet_leading_pt
d_pt_ratio_lep_tau	d_deltar_tau_lep	-	d_mass_transverse_met_l
p_met	d_pt_h	-	p_jet_leading_eta
p_lep_pt	p_tau_pt	-	d_mass_jet_jet
p_lep_eta	d_mass_jet_jet	-	p_jet_leading_phi
d_mass_MMC	p_met	d_mass_MMC	d_pt_h
——			
p_jet_leading_phi	d_prodeta_jet_jet	d_lep_eta_centrality	d_mass_vis
d_met_phi_centrality	p_jet_subleading_pt	p_jet_subleading_phi	p_jet_subleading_phi
p_tau_phi	p_lep_phi	p_tau_pt	d_prodeta_jet_jet
p_met_phi	d_mass_MMC	d_mass_vis	p_jet_subleading_pt
p_tau_eta	d_deltaeta_jet_jet	p_tau_phi	p_tau_pt
	——		
p_jet_subleading_pt	d_lep_eta_centrality	p_met_sumet	p_met_phi
d_pt_tot	p_jet_leading_phi	p_lep_pt	p_jet_all_pt
p_jet_subleading_eta	p_jet_leading_eta	p_lep_phi	p_met_sumet
d_prodeta_jet_jet	d_sum_pt	p_jet_subleading_pt	p_tau_eta
			——
p_jet_subleading_phi	p_jet_subleading_eta	p_jet_subleading_eta	p_met
p_met_sumet	p_tau_phi	d_pt_tot	d_met_phi_centrality
d_deltaeta_jet_jet	p_met_phi	p_jet_subleading_eta	d_pt_ratio_lep_tau
		——	
d_mass_jet_jet	p_jet_leading_pt	p_jet_subleading_phi	p_tau_phi
d_lep_eta_centrality	p_tau_eta	p_jet_subleading_pt	p_jet_subleading_eta
p_lep_phi	p_met_sumet	d_pt_h	p_lep_phi
p_jet_num	p_jet_num	p_met	d_pt_tot
d_pt_h	d_pt_tot	d_prodeta_jet_jet	p_lep_eta
d_sum_pt	p_lep_eta	p_lep_pt	d_sum_pt
p_jet_leading_pt	p_jet_subleading_phi	d_pt_tot	d_lep_eta_centrality
p_jet_all_pt	p_jet_all_pt	p_met_phi	p_lep_pt
p_jet_leading_eta	d_met_phi_centrality	p_lep_phi	d_deltar_tau_lep

3 Transformation der Eingabevariablen

3.1 Vorgehen

Gemäß der Aufgabenstellung in der Template-Datei, experimentierten wir mit den Transformationen „Decorrelate“ (D), „Gauß“ (G), und „Normalise“ (N). Um heraus zu finden, ob sie einen Einfluss auf das Ergebnis haben, wendeten wir sie in allen Kombinationen und Reihenfolgen auf den Input an.

3.2 Auswertung

Die Ergebnisse sind in Abbildung 2 gezeigt. Man sieht, dass es für die Fischer-Methode Sinn macht, die Daten zu dekorrelieren und zu Normaliseren. Die Reihenfolge hat keine Auswirkungen. Bei der BDT-Methode verschlechtern die Transformationen das Ergebnis. Außerdem gab es bei dieser Methode bei manchen Transforamtionen einen Programmfehler, den wir nicht weiter untersucht haben. In Abbildung 2 sind nur die erfolgreichen Abläufe der BDT-Methode gelistet und deshlab weniger als bei der Fischer-Methode.

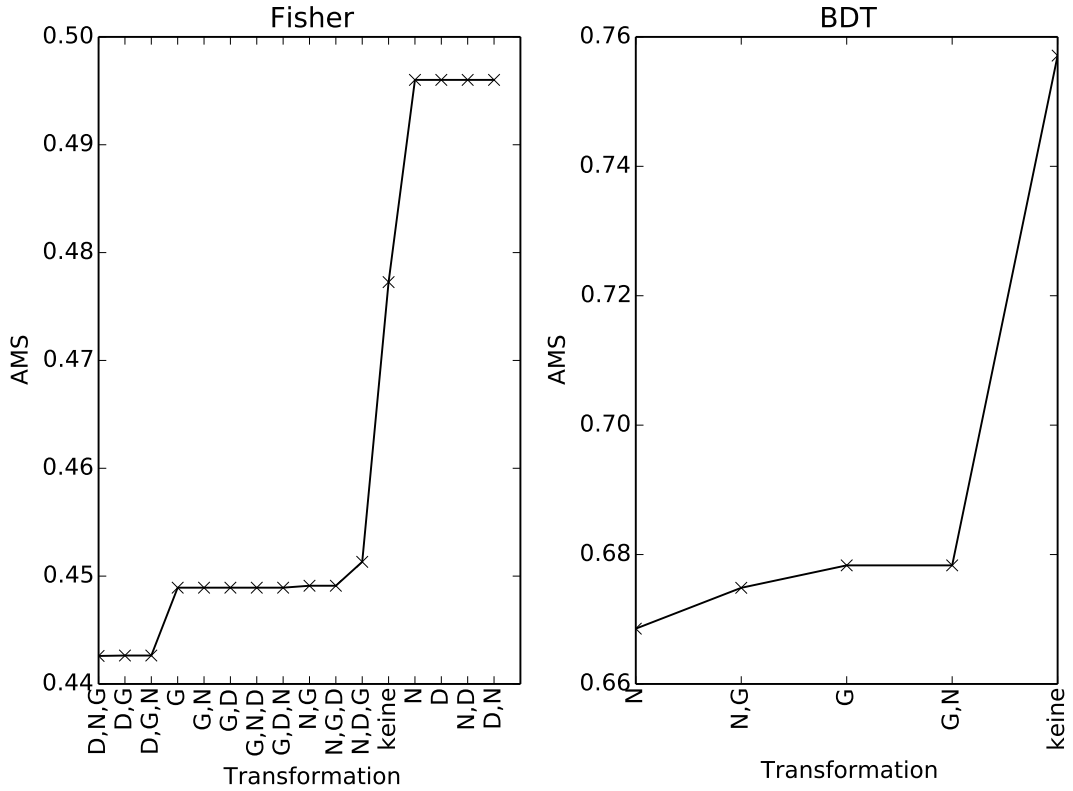

Abbildung 2: AMS über Transformation der Eingabevariablen.

4 Getting familiar with the project

4.1 Correlation of Variables

- determination of differences in correlations of signal and background
- removal of non relevant variables

4.2 Choosing a Classifier

5 improvement approach/Methodik

5.1 Improving the Classifier

5.2 Choosing the right cut

6 Conclusion