Datenanalyse

Heiko Rölke* und Marco Schmid

Allergiedaten Analysieren

SOSALL als Beispiel für die interdisziplinäre Zusammenarbeit im DAViS-Zentrum

https://doi.org/10.1515/iwp-2020-2120

Zusammenfassung: Das Zentrum für Datenanalyse, Visualisierung und Simulation (DAViS) der Partner FH Graubünden und Schweizer Institut für Allergie- und Asthmaforschung bietet Unterstützung in allen Bereichen der Forschung und Anwendung von Maschinellen Lernverfahren, Big Data, Visualisierung und Simulation. Im Artikel wird exemplarisch an einem Forschungsprojekt die Analyse komplexer Daten aus dem Gesundheitsbereich dargestellt.

Deskriptoren: Big Data, Datenanalyse, Deep Learning, Maschinelles Lernen, Simulation, FHGR, SIAF

Analyzing allergy data

SOSALL as an example for the interdisciplinary cooperation in the DAViS centre

Abstract: The Center for Data Analysis, Visualization, and Simulation (DAViS) at the University of Applied Science of the Grisons and the Swiss Center of Allergy and Asthma Research supports research on applications in topics like machine learning, big data, visualization, and simulation. The paper illustrates complex data analysis of life science data in a common research project.

Descriptors: Big Data, Data Analysis, Deep Learning, Machine Learning, Simulation, FHGR, SIAF

Analyser les données sur les allergies

SOSALL comme exemple pour la coopération interdisciplinaire au sein du centre DAViS

Résumé: Le Centre d'analyse, de visualisation et de simulation de données (DAViS) des partenaires FH Graubünden et l'Institut suisse de recherche sur les allergies et l'asthme offre un soutien dans tous les domaines de la recherche et de l'application des processus d'apprentissage automatique, du Big Data, de la visualisation et de la simulation. Dans l'article, l'analyse de données complexes issues du secteur de la santé est présentée à titre d'exemple dans un projet de recherche.

Descripteurs: Big Data, Analyse des données, Apprentissage approfondi, Apprentissage automatique, Simulation, FHGR, SIAF

Einführung

Deep Learning, Big Data und Simulationen auf Supercomputern – das sind nur einige der Themen, mit denen sich das neue Zentrum für Datenanalyse, Visualisierung und Simulation (DAViS) beschäftigt. Der Kanton Graubünden hat die Fachhochschule Graubünden (FHGR) und das Schweizerische Institut für Allergie- und Asthmaforschung (SIAF) beauftragt, Themen rund um Daten, Life Science und High-Performance Computing gemeinsam zu bearbeiten und anderen Forschenden und Industriepartnern mit Rat und Tat zur Seite zu stehen.

DAViS vereint drei inhaltliche Schwerpunkte mit drei Umsetzungsbereichen, die das Zentrum schon im Namen trägt: Datenanalyse, Visualisierung, Simulation. In diesen Schwerpunkten betreibt DAViS eigene Forschung, berät interne und externe Interessentinnen und Interessenten als Dienstleistung, bietet Infrastruktur an und beteiligt sich an der Lehre an der FH Graubünden und darüber hinaus.

In der Regel werden Forschungsprojekte gemeinsam mit externen Partnern angestoßen, können aber durch die interdisziplinäre Ausrichtung auch intern aufgesetzt werden. Alle durch DAViS abgedeckten Bereiche erfordern hohe Rechenleistung und Speicherkapazität. Dafür wird nach und nach eigene Hardware beschafft. Darüber hinaus ist DAViS eine Kooperation mit dem Schweizer Supercomputing-Center (CSCS) in Lugano eingegangen, so dass unter anderem auf den „Piz Daint" zugegriffen werden kann, den momentan sechst-schnellsten Rechner der Welt.

***Kontaktperson: Prof. Dr. Heiko Rölke,** Fachhochschule Graubünden, Schweizerisches Institut für Informationswissenschaft, Pulvermühlestrasse 57, 7000 Chur, Schweiz, E-Mail: heiko.roelke@fhgr.ch
Marco Schmid, BSc, Fachhochschule Graubünden, Schweizerisches Institut für Informationswissenschaft, Pulvermühlestrasse 57, 7000 Chur, Schweiz, E-Mail: marco.schmid@fhgr.ch

Tabelle 1: Ausschnitt der Analysen zur Zielvariable ‹diagnosis_location›.

Feature/Variable	Anzahl Beobachtungen	Testname	p-Wert	Gruppen-Mittelwerte	post-hoc-Test
eczema_ever	[60, 56, 52, 49]	chi-square	1.86E-43		
medication_steroidcreams	[60, 54, 52, 49]	chi-square	1.48E-40		
farmanimal_contact_child	[60, 53, 52, 49]	chi-square	1.91E-33		
farmanimal_contact_mother	[60, 53, 52, 48]	chi-square	4.23E-32		
fuel_cooking_Electricity_Gas	[60, 56, 52, 49]	chi-square	2.62E-30		
medication_antihistamines	[60, 54, 52, 49]	chi-square	2.94E-25		
sunlight_exp_winter	[60, 52, 50, 48]	kruskal	7.64E-21	AD_Rural=3.138 / AD_Urban=0.871 / HC_Rural=4.269 / HC_Urban=0.977	Dunn
sp_any	[55, 55, 52, 49]	chi-square	3.65E-20		
fuel_cooking_Paraffin Stove	[60, 56, 52, 49]	chi-square	1.89E-19		

Die Datenanalyse mittels maschineller Lernverfahren ist in den vergangenen Jahren von einer Nischenanwendung zu einer wichtigen Methode in zahlreichen Anwendungsfeldern gereift. Vor allem das Feld des „Deep Learning“, also Lernen mittels künstlicher neuronaler Netze, wird in immer mehr Bereichen eingesetzt. Das erfordert allerdings sowohl Fachwissen als auch ausreichend Rechenleistung für das Training der Algorithmen. Dieses Hindernis war einer der Gründe für die Gründung und Förderung von DAViS: Im Zentrum wird sowohl die Expertise gebündelt als auch die Infrastruktur aufgebaut, um maschinelles Lernen erfolgreich in die Praxis zu bringen.

Ein Fallbeispiel

Ein Beispiel für ein internes Datenanalyse-Projekt zwischen SIAF und FH Graubünden ist „MLM-SOS-ALL“, in dem mit Machine Learning und Modelling nach molekularen, genetischen und umweltbedingten Faktoren gesucht wird, die für die Entstehung und Verbreitung allergischer Krankheiten verantwortlich sind. Die zugrundeliegenden Daten wurden vorgängig in der SOS-ALL Studie (South-African – Swiss: Mechanisms of the Development of Allergy) in einem Konsortium aus Schweizerischem Institut für Auslandsforschung (SIAF), Universität Kapstadt, Kinderspital Zürich und Dermatologischer Klinik des Universitätsspitals Zürich erhoben und bestehen aus einem großen RNA-Sequenzier-Datensatz und detaillierten Angaben zu den Patienten, ihren Lebensumständen und der Krankengeschichte. Die Probanden in der SOS-ALL Studie sind Kinder aus Stadt und Land, aus der Schweiz und aus Südafrika, mit und ohne atopischer Dermatitis.

Die umfangreiche Datenanalyse wird gemeinsam von den DAViS-Partnern SIAF und FH Graubünden vorangetrieben. In einer kombinierten Analyse der Datensätze, die über die bisher verwendeten biostatistischen Methoden hinausgeht, sollen Hinweise gefunden werden, die zu einem besseren Verständnis der komplexen Zusammenhänge führen, die das Auftreten allergischer Erkrankungen vor allem in der Stadt fördern. Zudem sollen Risikofaktoren und Biomarker für die Entstehung von Allergien identifiziert werden, die zu Präventions-Maßnahmen und verbesserter Diagnostik genutzt werden können.

Die Analyse ist so aufgebaut, dass Fragebogen und RNA-Daten zuerst getrennt aufbereitet und analysiert werden und die Ergebnisse dann anhand der pseudonymisierten IDs der Probanden zusammengeführt werden. Die Analyse läuft derzeit noch, so dass noch keine abschließenden Ergebnisse genannt werden können.

Die Fragebogenanalyse verwendet sowohl einen „klassischen“ Ansatz mit statistischen Tests als auch einen Ansatz mit Machine Learning. Der Fragebogen umfasst die Daten von 210 Probanden. In einem ersten Schritt haben wir uns einen Überblick verschafft, unklare Bezeichner abgeklärt und vereinheitlicht, mehrfach verwendete Datenfelder aufgeteilt usw. Der Datensatz weist einen Fehlbestand („missing values“) von ca. 29 Prozent auf. Nur ein Teil davon ist strukturell bedingt, einige Datenspalten lassen sich aufgrund des hohen Fehlbestands nicht nutzen. Die Datenvorbereitung („data pre-processing“) ergibt einen Datensatz von knapp 20.000 Datenfeldern.

Aus dem so vorbereiteten Datensatz werden für die statistische Analyse zuerst Zielvariablen anhand der Variablen im Versuchsaufbau ausgesucht, die mit den Wer-

ten im Fragebogen korrelieren können. In unserem Fall sind dies die Werte „Diagnose“ (also das Vorliegen einer atopischen Dermatitis oder nicht), „Wohnort“ (Stadt oder Land) und die Verbindung aus den beiden. Tabelle 1 zeigt für einen kleinen Ausschnitt ausgewählter Werte die Ergebnisse der Analyse. Die tatsächliche Tabelle ist in beiden Dimensionen wesentliche umfangreicher.

Eine solche Tabelle mit allen Variablen und Analysewerten wird jedoch schnell unübersichtlich, so dass alle Ergebnisse mit numerischen Werten auch graphisch dargestellt werden. Die folgende Graphik in Abbildung 1 zeigt beispielsweise das Feature (eine Variable) „log_blood_count_monocytes“, also der Logarithmus eines spezifischen Blutwertes und als Zielvariable „diagnosis_location“, also die Kombinationen aus Diagnose und Wohnort, inklusive der Post-Hoc-Test-Ergebnisse. Dabei stehen schwarze-gestrichelte Linien für signifikante Gruppenunterschiede ($p < 0.05$) und schwarze durchgezogene Linien für hoch signifikante Gruppenwertunterschiede ($p < 0.01$). Anhand von Graphiken wie der in Abbildung 1 lassen sich die Zwischenergebnisse der Analyse gut mit den Domänenexperten besprechen und bewerten.

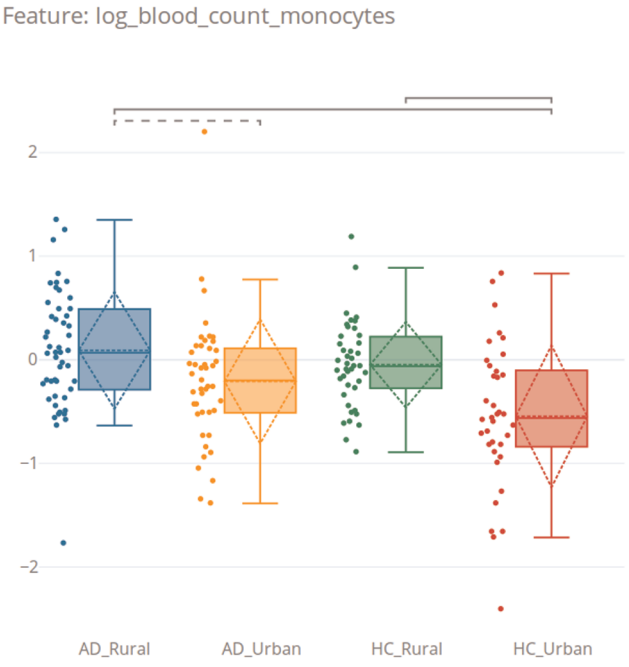

Abbildung 1: Gegenüberstellung von Analysewerten.

Klassische statistische Analysen lassen sich schnell und einfach durchführen und führen im Idealfall auch schon zu verwertbaren Ergebnissen. Ungünstig ist jedoch gerade bei großen Datensätzen die hohe Anzahl an Tests und vor allem bei komplexen Zusammenhängen die Grundannahme der Unabhängigkeit einzelner Faktoren voneinander. Ein Test auf mehrere miteinander gekoppelte Abhängigkeiten führt aber schon bei einem überschaubaren Datensatz wie dem vorliegenden zu einer sehr hohen Anzahl von Kombinationsmöglichkeiten – exponentiell auf der Anzahl der Variablen. Dies macht es schwierig, die Übersicht zu behalten und erfolgsversprechende Analysewege zu entdecken. Abhilfe schaffen kann hier das Machine Learning, das Teile des Suchens und Ausprobierens automatisiert.

Aus der großen Auswahl an Algorithmen im maschinellen Lernen kommen hierfür vor allem die überwachten Lernalgorithmen (supervised learning) in Frage – zu Hintergründen siehe beispielsweise Igual und Segui (2017). Vor der Anwendung sind aber noch weitere Datenkodierungen notwendig, zum Beispiel um nicht-numerische in numerische Werte umzuwandeln. Dazu wird nach einem festen Übertragungsschema jedem Wert eine Zahl eindeutig zugeordnet, so dass sich später die ursprünglichen Werte wieder rekonstruieren lassen. Zusätzlich müssen fehlende Werte ergänzt (imputiert) werden – oder die entsprechenden Variablen müssen ausgelassen werden.

Im maschinellen Lernen wird im Regelfalle so verfahren, dass die Daten in einen Trainings- und einen Testanteil aufgeteilt werden. Anhand des Trainingsanteils wird der Lernalgorithmus trainiert und anhand des Testteils der Erfolg des Trainings überprüft. Die Trainingsdaten werden nochmals aufgeteilt in die Daten für das eigentliche Training und einen Validierungsdatensatz. Diese Aufteilung wird üblicherweise stetig verändert und wiederholt, um eine Verfälschung durch eine zufällig ungünstige Aufteilung zu verhindern. Auch diese Arbeitsschritte werden automatisiert und oft wiederholt. Die Aufteilung des Datensatzes ist noch unabhängig vom eigentlichen Lernvorgang.

Ziel der maschinellen Lernaktivität ist ein speziell auf die Daten angepasstes Verfahren, das sowohl für die Trainingsdaten als auch für die Testdaten gute Ergebnisse liefert. Ein häufig auftretendes Problem, gerade bei kleineren Datensätzen ist die Überanpassung (overfit), also das „Auswendiglernen“ des Trainingsdatensatzes. In diesem Fall liefert die Analyse mit den Testdaten schlechtere Werte. Abbildung 2 zeigt einen typischen Fall: Während die Trainingsdaten sehr gute Werte liefern, sind die Ergebnisse im Testfall nicht so gut, wenn auch noch brauchbar.

In Abbildung 2 kann man schon gut sehen, wie mit steigender Anzahl an ausgewählten Variablen (Features) die Vorhersage besser wird und schließlich ein Maximum erreicht oder jedenfalls einen Wert, der mit zusätzlichen Variablen nicht mehr stark ansteigt. Es stellt sich die Frage, welche Variablen den höchsten Anstieg verursachen und welche eher ausgelassen werden können. Bei der Be-

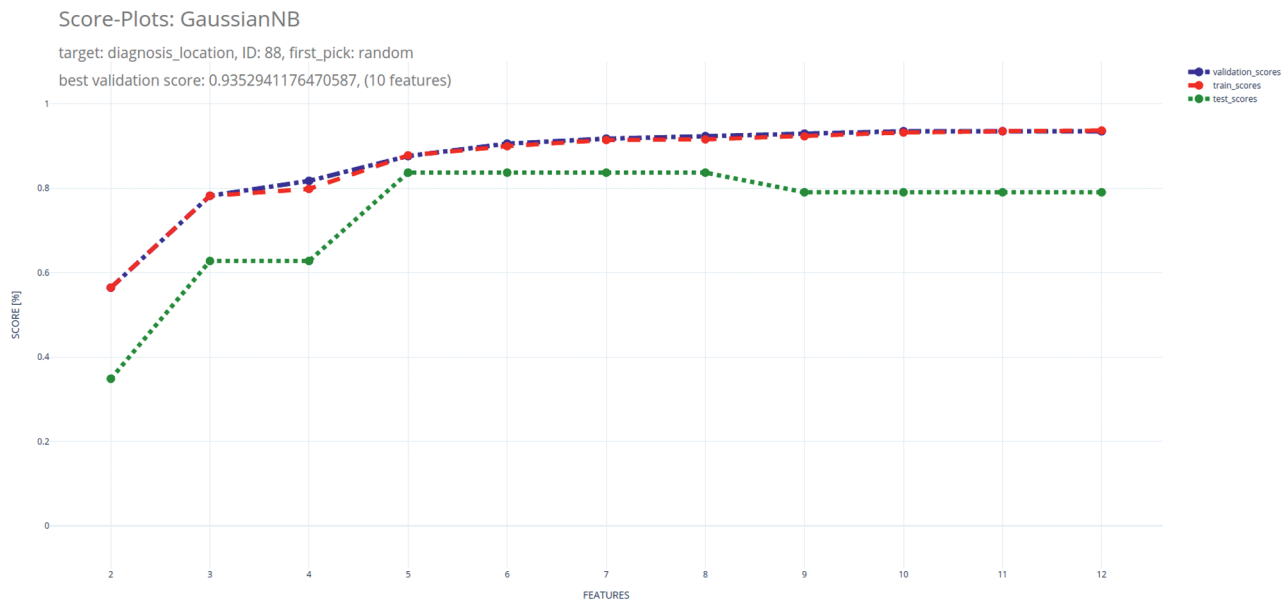

Abbildung 2: Trainingsdaten und Testdaten.

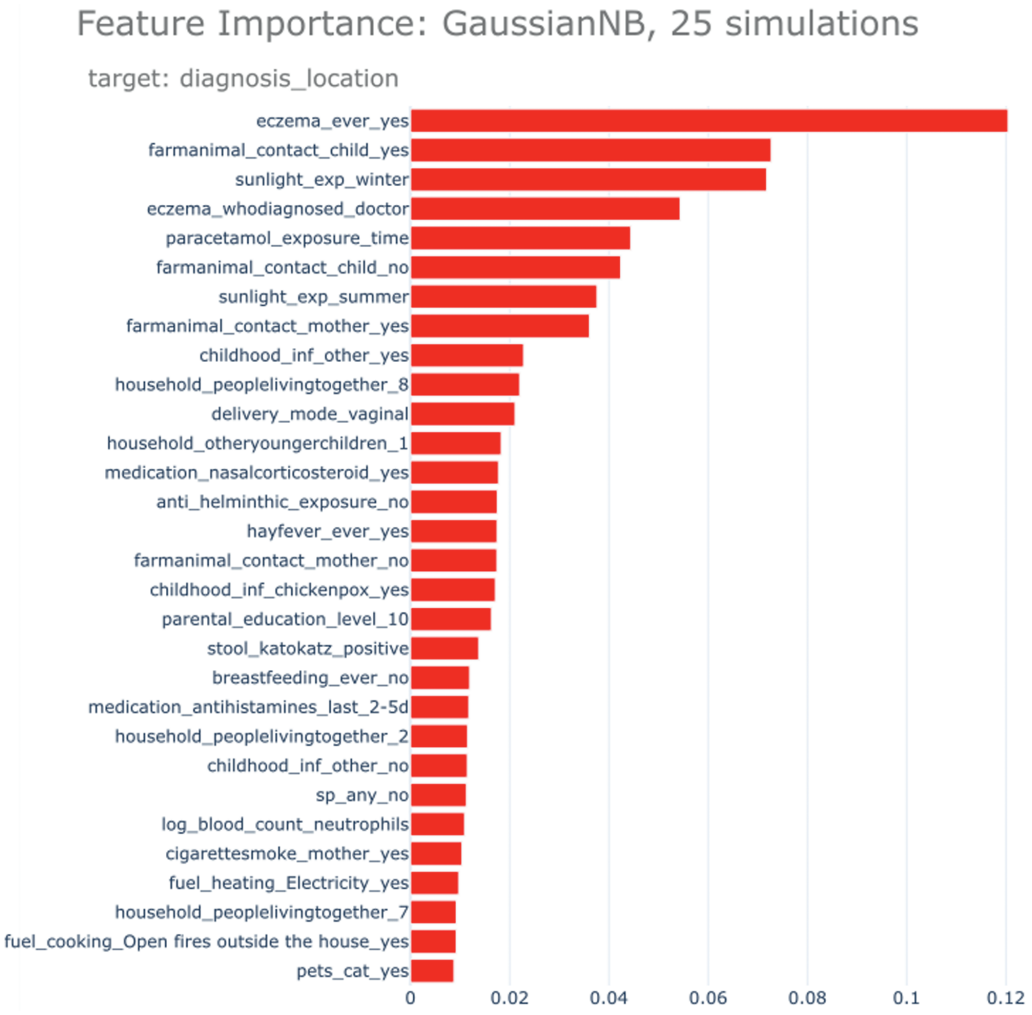

Abbildung 3: Vergleich der besten Prädiktoren.

rechnung der ‹feature importance›, also der Wichtigkeit der Variablen, werden iterativ Kombinationen von Variablen ausprobiert (stepwise forward selection) und pro Simulation die jeweils beste Kombination als Prädiktor verwendet. Die Simulation wird nacheinander repetitiv auf unterschiedlichen Trainings- und Testdaten ausgeführt. Dabei wird registriert, welche Variablen pro Simulation früh (oder spät) gewählt werden.

Abbildung 3 zeigt eine Zusammenfassung aller verwendeten Features für die prädiktiven Modelle nach einer vorgegebenen Anzahl Simulationen. Hieraus kann abgeleitet werden, welche Variablen oft für die Erstellung eines möglichst guten Modells herangezogen werden. Dies sind teilweise vielversprechende Objekte für weitere Analysen, teilweise sieht man schon als Laie, dass einige der Variablen nicht für die Analyse geeignet sind, zum Beispiel gleich die Erste (eczema_ever_yes), die fast der eigentlichen Diagnose entspricht. Andere Variablen sind vielversprechender, wie der Kontakt mit landwirtschaftlichen Tieren als Kind, die durchschnittliche Zeit, die draußen verbracht wird, und einige weitere.

Fazit

Wie schon in der Einleitung geschrieben, laufen die Analysen derzeit noch, so dass an dieser Stelle noch keine Ergebnisse genannt werden können. Eine wichtige Lehre war die Bedeutung der engen Zusammenarbeit mit Domänenexperten, um die Analyse laufend neu ausrichten zu können. Einige zuerst gefundene Ergebnisse stellten sich als sinnlos heraus, da die Variablen nicht voneinander unabhängig waren. Durch die enge Kooperation im DAViS-Zentrum mit den Experten für Life Science am SIAF wurde dies schnell entdeckt und behoben. Hilfreich für eine gute Zusammenarbeit über die Disziplingrenzen hinweg ist die intuitive Visualisierung der Ergebnisse. Dadurch werden eine Grundlage für die schnelle Erfassung der Ergebnisse gelegt und Diskussionen ermöglicht. Visualisierungen spielen nicht nur im DAViS-Zentrum eine wichtige Rolle, sondern werden zukünftig auch im Bachelor und insbesondere im Master-Studium der Informationswissenschaft an der FH Graubünden eine tragende Rolle spielen.

Literatur

Igual, L., Seguí, S. (2017). Introduction to Data Science, Springer-Verlag, DOI 10.1007/978-3-319-50017-1.

Prof. Dr. Heiko Rölke
Fachhochschule Graubünden
Schweizerisches Institut für Informationswissenschaft
Pulvermühlestrasse 57
7000 Chur
Schweiz
heiko.roelke@fhgr.ch

Prof. Dr. Heiko Rölke wurde an der Universität Hamburg in Informatik promoviert und ist seit 2017 Dozent für Data Science an der FH Graubünden. Seine Schwerpunkte liegen in der Modellierung und Analyse komplexer, verteilter und nebenläufiger Systeme. Seine Forschungsinteressen liegen im Bereich der Modellierung, Implementierung und insbesondere Analyse von verteilten Systemen – speziell Multiagentensystemen – und formalen Modellierungs- und Analysetechniken.

Marco Schmid, BSc
Fachhochschule Graubünden
Schweizerisches Institut für Informationswissenschaft
Pulvermühlestrasse 57
7000 Chur
Schweiz
marco.schmid@fhgr.ch

Marco Schmid studierte Sport an der Universität Basel und nahm nach einigen Jahren in der Sportwissenschaftlichen Forschung ein Zweitstudium an der ZHAW Zürich in Umweltingenieurwesen auf, in dem er sich mit der Datenanalyse beschäftigte. Nach seinem Bachelor arbeitete er in der Privatwirtschaft als Data Scientist und Entwicklungsingenieur. Seit Juli 2019 ist Marco Schmidt wissenschaftlicher Mitarbeiter im Schweizerischen Institut für Informationswissenschaft (SII).