

Zusammenfassung

Johannes Alois Vey
Dr. sc. hum.

Comparison of Regression and Machine Learning Methods for Variable Selection to Develop a Clinical Prediction Model

Fach/Einrichtung: Medizinische Biometrie
Doktorvater: Prof. Dr. sc. hum. Meinhard Kieser

Klinische Prädiktionsmodelle (KPM) bekommen in der Gesundheitsversorgung eine immer größer werdende Bedeutung, da sie die Vorhersage einer klinischen Zielgröße ermöglichen und somit die Entscheidungsfindung eines Arztes/einer Ärztin und eines Patienten/einer Patientin unterstützen können. Eine häufige Herausforderung bei der Entwicklung eines KPMs ist die Identifizierung von wichtigen Prädiktorvariablen. Für die Variablenselektion und Entwicklung eines KPMs gibt es eine Vielzahl an unterschiedlichen Methoden: Während einige auf traditionellen Regressionsmodellen basieren, gehören andere dem Bereich des maschinellen Lernens an. Ziel dieser Thesis war es daher, Methoden unterschiedlicher Modellierungstechniken für die Entwicklung eines KPMs einschließlich Variablenselektion für eine stetige Zielgröße in niedrig-dimensionalen Daten zu vergleichen, um daraus Empfehlungen für die praktische Anwendung abzuleiten.

Im ersten Schritt wurde zunächst ein Simulationsdesign entwickelt, um einen möglichst fairen Vergleich sowie die Ausarbeitung von Stärken und Schwächen der verschiedenen Methoden zu ermöglichen. Hierzu wurden vier datengenerierende Prozesse mit steigender Komplexität erstellt, die realistische Datenstrukturen mit praxisrelevanten Herausforderungen generierten. Anhand dieser simulierten Daten wurden anschließend etablierte und häufig angewandte Methoden verglichen: lineare Regression mit stufenweiser Selektion (LMSS), regularisierte lineare Regression mit dem Elastic Net Strafterm (ENET), Gradient Boosting mit jeweils linearen Regressionsmodellen (GBM) und Entscheidungsbäumen (GBT) als Basis-Lerner sowie die Boruta (RFB) und Hapfelmeier (RFH) Methoden für Random Forest. Zudem wurden Multivariable Fractional Polynomial (MFP) Regressionsmodelle als Bezugsmethode angewandt.

Alle Methoden selektierten eine zunehmende Zahl an Variablen mit steigender Fallzahl des Datensatzes. Während LMSS, RFB und RFH bessere Ergebnisse hinsichtlich des korrekten Einschlusses von wichtigen Prädiktorvariablen und des korrekten Ausschlusses von irrelevanten Variablen zeigten, resultierten ENET, GBM und GBT in Modellen, die nahezu alle Variablen einschlossen. In den Szenarien mit niedriger Komplexität demonstrierte LMSS die besten Selektionseigenschaften und identifizierte auch Prädiktoren mit nicht-linearer funktionaler Form zur Zielgröße. In den komplexeren Szenarien sank die Einschlusshäufigkeit für nicht-lineare Prädiktoren, die nicht-lineare Zusammenhänge mit irrelevanten Variablen aufwiesen, insbesondere für LMSS. In den komplexesten Szenarien zeigten RFB und RFH die überlegensten Selektionseigenschaften. Hinsichtlich der prädiktiven Genauigkeit in Testdatensätzen erzielten alle Methoden vergleichbare Resultate, wobei sich die Genauigkeit der entwickelten KPMs grundsätzlich mit steigender Fallzahl verbesserte. Während ENET, GBM und GBT aufgrund von Regularisierung eine sehr gute Kalibrierung zeigten, neigten RFB und RFH zu Unter- und LMSS zu Überanpassung an die Trainingsdaten in gewissen Szenarien. Ergänzend zur Simulationsstudie wurden die Methoden auf einen realen Datensatz angewandt, um KPMs für den intraoperativen Blutverlust während Lebertransplantationen zu entwickeln. Das entwickelte Simulationsdesign ist frei verfügbar und kann für weiterführende Forschungen, wie beispielsweise der Untersuchung zusätzlicher Methoden, verwendet werden. Darüber hinaus kann das Design um weitere Aspekte, wie z.B. eine größere Anzahl an Variablen, ergänzt werden.

Diese Thesis bietet einen umfangreichen Vergleich von traditionellen Regressionsmethoden und Methoden des maschinellen Lernens für Variablenselektion und der Entwicklung eines KPMs für eine stetige Zielgröße in niedrig-dimensionalen Daten. Für alle Methoden ist eine Fallzahl von 250-500

Beobachtungen nötig, um Prädiktorvariablen mit einer ausreichenden Sicherheit zu identifizieren und eine adäquate Prädiktionsgenauigkeit zu erzielen. LMSS kann für die Anwendung in Daten mit niedriger Komplexität sowie, ggf. mit Anpassungen, in komplexeren Datenstrukturen empfohlen werden. RFB und RFH können insbesondere für komplexere Daten, vorzugsweise wenn Interaktionen zwischen Variablen vorliegen können, empfohlen werden.