MedWiss.Online

Einsatz von Random Forests in der Präzisionsmedizin

Die Zielsetzung der Präzisionsmedizin ist es, jeden Patienten zielgerichtet und maßgeschneidert therapieren zu können

Dies erfordert die Translation vieldimensionaler Daten in die klinische Praxis, welche maßgeblich durch künstliche Intelligenz unterstützt werden kann, indem Methoden des maschinellen Lernens wie beispielsweise Zufallswälder (engl. Random Forests) eingesetzt werden.

Prof. Inke König und ihr Team vom Institut für medizinische Biometrie und Statistik beschäftigen sich im entsprechenden Forschungsprojekt intensiv mit diesem mathematischen Verfahren und adressieren dabei die folgenden medizinischen Fragestellungen:

1. **Welche neuen diagnostischen Taxonomien entstehen, wenn wir Patienten aufgrund genomischer Daten mit nicht-überwachten Zufallswäldern klassifizieren?**
   In diesem Teilprojekt werden Algorithmen zum nicht-überwachten Lernen durch Zufallswälder modifiziert, an vorliegende Datenstrukturen angepasst und in Simulationsstudien miteinander verglichen. Die Algorithmen werden u.a. für die Analyse genetischer Ähnlichkeiten in einem Projekt des Deutschen Zentrums für Herz-Kreislauf-Forschung (DZHK) zu koronaren Herzerkrankung und verwandten Phäntoypen verwendet.
2. **Wie lassen sich Risikoscores aus Zufallswäldern inhaltlich interpretieren?**
   Es werden Maße erweitert, die die Wichtigkeit einzelner Variablen für die Klassifikation in Zufallswäldern schätzen. Des Weiteren ermöglicht die Darstellung von Zufallswäldern durch repräsentative Bäume eine vereinfachte Interpretierbarkeit eines Scores. Neben Simulationsstudien zur Untersuchung der methodischen Eigenschaften repräsentativer Bäume werden die Methoden auf Daten der DFG-Forschergruppe 2488 angewendet.
3. **Wie können zur Integration verschiedener Omics-Daten Informationen über funktionelle und strukturelle Zusammenhänge in Zufallswäldern integriert werden?**
   Detaillierte Informationen über Zusammenhänge zwischen Genen oder anderen Molekülen sind in Form von Netzwerken öffentlich verfügbar. Es wird untersucht, in welcher Form dieses Wissen in den Trainingsprozess von Zufallswäldern integriert werden kann. In Simulationen werden konzeptuelle, vereinfachte Situationen modelliert, aber auch Szenarien, die stark an experimentellen Daten orientiert sind.
4. **Wie können Zufallswälder bei longitudinalen Daten oder abhängigen Beobachtungen verwendet werden?**
   Umfangreiche Simulationsstudien werden durchgeführt, um verschiedene Ansätze zur Analyse longitudinaler und/oder abhängiger Daten systematisch zu vergleichen und Empfehlungen für konkrete Anwendungsszenarien geben zu können. Eine Anwendung geeigneter Methoden ist auf Daten der ALLIANCE-Kohorte des Deutschen Zentrums für Lungenforschung (DZL) geplant, um die Entwicklung von Asthma vorherzusagen.

Falls Sie Fragen zum Projekt oder verwandten Themen haben, besuchen Sie doch einfach die Webseite des IMBS www.imbs.uni-luebeck.de
oder werfen Sie eine Blick in die entsprechenden Veröffentlichungen:

- Abegaz F, Chaichoompu K, Génin E, Fardo DW, König IR, Mahachie John JM, and Steen KV.

(2019) Principals about principal components in statistical genetics. Brief Bioinform 20: 2200-16
- Boulesteix A-L, Wright MN, Hoffmann S, and König IR. (in press) Statistical learning approaches in the genetic epidemiology of complex diseases. Hum Genet
- Degenhardt F, Seifert S, Szymczak S. (2019) Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform. 2019 20: 492-503
- Gola D, Erdmann J, Müller-Myhsok B, Schunkert H, and König IR. (2020) Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. Genet Epidemiol 44: 125-38
- König IR, Fuchs O, Hansen G, von Mutius E, and Kopp M. (2017) What is Precision Medicine? Eur Respir J 50: 1700391
- Nembrini S, König IR, and Wright MN. (2018) The revival of the Gini Importance? Bioinformatics 34: 3711-8
- Seifert S, Gundlach S, Szymczak S. (2019) Surrogate minimal depth as an importance measure for variables in random forests. Bioinformatics 35: 3663-3671
- Seifert S, Gundlach S, Junge O, Szymczak S. (2020) Integrating biological knowledge and gene expression data using pathway guided random forests: A benchmarking study.Bioinformatics. doi: 10.1093/bioinformatics/btaa483. Online ahead of print.