Ruprecht-Karls-Universität Heidelberg

Medizinische Fakultät Mannheim

Dissertations-Kurzfassung

Hierarchische Kombination von Klassifikationsverfahren zur Analyse von Genexpressionsdaten

Autor: Jasmin Müller
Institut / Klinik: Zentrum für Medizinische Forschung Mannheim
Deutsches Krebsforschungszentrum Heidelberg (DKFZ)
Doktorvater: Prof. Dr. N. Gretz

Einführung und Motivation

In der Krebstherapie werden nach wie vor Therapien mit starken Nebenwirkungen angewendet, wie zum Beispiel die systemische Therapie mit cytotoxischen Substanzen. Diese aggressiven Verfahren sind zur Zerstörung der Tumormasse notwendig. Die Behandlung für den einzelnen Patienten kann schonender gestaltet werden, indem man die Patienten in Untergruppen einteilt und auf Grund dieser Unterteilung unterschiedliche, risikoadaptierte Therapien anwendet. Die Bedeutung molekulargenetischer Daten, wie zum Beispiel Genexpressionsprofilen, für die Individualisierung der Therapie konnte in einer Reihe von Studien gezeigt werden.

Die Auswertung solcher hochdimensionaler Daten wirft allerdings statistische Probleme auf, da zu einem einzigen Patienten oft mehr als 1000 Merkmale erhoben werden, aber meist nur bei wenigen Patienten die entsprechende Krebserkrankung auftritt. Maschinelle Lernverfahren ermöglichen es, solche Daten zu nutzen.

Diese Verfahren gehen von Trainingsdaten mit bekannter Klasseneinteilung aus und können neue Fälle entsprechend einer 'erlernten' Verallgemeinerung klassifizieren. Dazu wird der Klassifikator zunächst auf Daten mit bekannter Klassifikation trainiert, um dann die verschiedenen Klassen mittels der erfassten Merkmale zu unterscheiden.

Es gibt mehrere Möglichkeiten, einen solchen Klassifikator zu realisieren. Ein Verfahren zur Klassifikation aus dem Bereich der maschinellen Lernverfahren sind Support-Vektor-Maschinen (SVM), die wiederholt erfolgreich auf Genexpressionsdaten angewendet werden konnten.

SVMs wurden entwickelt, um Fragestellungen mit zwei verschiedenen Klassen zu bearbeiten. Da in dieser Arbeit zwischen mehr als zwei Unterarten eines Tumors differenziert werden soll, ist es notwendig, das Verfahren auf die Mehrklassenproblematik anzupassen. Kombiniert man SVMs mit binären Entscheidungsbäumen, erhält man ein Dendrogramm der Klassen, das die Reihenfolge wiedergibt, in der diese Klassen durch SVM-basierte Diskriminanzanalyse voneinander unterschieden werden. Diese Darstellungsweise entspricht den häufig verwendeten Clusteranalysen von Genexpressionsdaten, beruht jedoch auf einer neuartigen Berechnungsweise und bietet einen anderen Blickwinkel auf die Patientendaten.

Ergebnisse

Die Ergebnisse dieser Arbeit auf realen Datensätzen unterschiedlicher Tumorarten belegen, dass die Klassifikationsgenauigkeit des SVM-Binärbaums mit etablierten Verfahren zur Multiklassen Klassifikation mittels einer Support-Vektor-Maschine vergleichbar ist.

Die Topologien, die mit dem in dieser Arbeit entwickelten Verfahren ermittelt wurden, zeigen Parallelen zu bereits bekannten biologischen Zusammenhängen der verschiedenen Klassen. Es besteht somit die Möglichkeit, auf anderen Datensätzen neue Zusammenhänge zwischen Klassen zu ermitteln, die dann beispielsweise mittels gezielter Laborversuche weiter analysiert werden können. Dadurch könnten sich neue Ansätze im Bereich der Systembiologie sowie der Modellierung von Tumorentwicklungen ergeben.

Zusammen gefasst ermöglicht das entwickelte Verfahren eine dem Stand der Technik entsprechende Klassifikation von Multiklassenproblemen sowie eine neuartige Analyse von grundlegenden Zusammenhängen der Klassen und ihre graphische Darstellung.