Simple and Effective Visual Models for Gene Expression Cancer Diagnostics

Handout

- Thema: Trennen von Krebs-Genexpressionsdaten in 2D-Graphen mit Hilfe des VizRank-Algorithmus
- Quelle der Daten: Microarrays mit den Expressionswerten karzinogener Zellen
- Jeder Datensatz besteht aus mehreren tausend Genen und nur ein paar Patienten
- Scatterplot- und Radviz-Projektionen werden zum Darstellen der Daten benutzt
- Ziel ist es, die Krebsarten (Klassen) deutlich zu trennen
- Gesucht werden also Darstellungen, die diese Eigenschaft haben
- Auf Grund der großen Menge an möglichen Darstellungen wird eine heuristische Suche (RELIEF-Summen) benutzt, um viel versprechende Darstellungen zu filtern
- Die gefilterten Darstellungen werden mit dem VizRank-Algorithmus genauer bewertet
 1. Anlernen der Daten mit dem k-nearest-neighbour Algorithmus und euklidischer Distanz
 2. Bewerten der Vorhersagegenauigkeit mit der leave-one-out cross-validation
 3. An Hand der Bewertung ein Ranking machen
- Die Bewertung ist ein Indiz für die Auftrennung der Datenpunkte in der Darstellung
- Mit Hilfe der Darstellungen lassen sich Krebsarten unterscheiden und bestimmen.
- Zusätzlich sind die Gene, die an den besten Darstellungen beteiligt sind, oft spezifisch für die jeweilige Krebsart