Übungen Musterklassifikation

Wintersemester 2002/2003

Prof. Dr. Stefan Posch
Dipl.Inform. Denis Williams
Dipl.Inform. Gregor Erz

Institut für Informatik
Universität Halle

Blatt 2

Eines der aktuellen Forschungsgebiete in der Biologie ist die Analyse von Genexpressionen. Ausgehend von dem Wissen, dass verschiedene Krebsarten mit der Bildung verschiedener Proteine zusammenhängen - einerseits solche, die für den Krebs verantwortlich sind, andererseits solche, die ihn bekämpfen, ist man daran interessiert herauszufinden, welche Proteine verstärkt/vermindert (die Genexpression ist also erhöht oder verringert) gebildet werden.

Dazu wird präparierte und fluoreszierend markierte mRNA aus Gewebsproben auf ein sogenanntes Microarray aufgebracht, das komplementäre Basenfolgen der zu untersuchenden Proteine enthält. Die Proteine der Gewebsprobe binden sich an die zugehörigen Basen - je nach Anzahl der gebundenen Proteine erscheint dieser Ausschnitt auf dem Microarray durch die Fluoreszenz heller oder dunkler. Der Chip wird eingescannt und der Helligkeitswert jedes Proteins wird in einen Zahlenwert umgewandelt.

In diesem Falle wurden mehrere Knochenmark-Proben mit je etwa 35403 Genen (jede Probe ein Chip) von Patienten mit Leukämie (ALL) oder Ewing (Knochenmark-Krebs), sowie gesundes Knochenmark (KM) eingescannt. Zur Analyse der Daten gehört auch die Klassifikation der Proben, bzw. Zuordnung von noch nicht klassifizierten Proben zu einer der Gruppen, oder das Auffinden von Untergruppen und Ausreissern.

Nun sollen auf die Daten dieser drei Arten von Gewebsproben verschiedene Klassifikationsverfahren angewendet, also möglichst ähnliche Proben bzw. im günstigsten Fall die jeweiligen Knochenmark-Proben zusammen in ein Cluster gebracht werden. Dabei ist zu beachten, dass man gerade bei der Verwendung von zufälligen Anfangspunkten völlig verschiedene Ergebnisse erhalten kann. Für Tests ist also auch interessant, zwischen welchen Werten der Quantisierungsfehler schwankt. Zur Vereinfachung wird nur eine Auswahl von Genen betrachtet.

Die Daten sind, neben den Übungsblättern, zu finden unter:

http://www.informatik.uni-halle.de/~posch/LEHRE/MK02/material.html

- in GenDaten.txt die bereits klassifizierten Proben, und in ProbenXYZ.txt 2 noch unbekannten Proben.

Aufgabe 2.1 Ermitteln Sie für die Unterscheidung zwischen den Probensorten nützliche Merkmale, d.h. Gene. Bedenken Sie dabei, dass die tatsächliche Anzahl der Gene mit 35403 erheblich höher ist, als die Zusammenstellung von Genen für diese Aufgaben

- eine Auslese per Hand ist also nicht möglich. Schreiben Sie ein Programm, das solche Merkmale findet und aussortiert.

Aufgabe 2.2 Programmieren Sie den k-means-Algorithmus mit mindestens 4 verschiedenen Distanzfunktionen. Testen Sie Ihr Programm sowohl mit den vorgebenen Genen als auch mit den aus Aufgabe 1 aussortierten. Welche Cluster werden gebildet und welche Proben werden korrekt bzw. falsch zugeordnet? Zu welchen Clustern werden die zwei unbekannten Proben am ehesten zugeordnet? Geben Sie jeweils die Quantisierungsfehler an.

Aufgabe 2.3 Verbinden Sie den k-means aus Aufgabe 2 mit dem LBG-Algorithmus. Testen Sie Ihr Programm.

Aufgabe 2.4 Verbinden Sie den k-means/LBG aus Aufgabe 3 mit dem LVQ3-Algorithmus. Testen Sie Ihr Programm.

Aufgabe 2.5 Vergleichen Sie die Ergebnisse mit denen, die bei rein zufälliger Initialisierung der Gebietszentren zustande kommen.

Aufgabe 2.6 Programmieren Sie einen Nearest-Neighbour-Klassifikator mit mind. 2 verschiedenen Distanzfunktionen. Testen Sie Ihr Programm.

Aufgabe 2.7 Programmieren Sie einen m-Nearest-Neighbour-Klassifikator mit mind. 2 verschiedenen Distanzfunktionen. Testen Sie Ihr Programm.

Aufgabe 2.8 Programmieren Sie einen Hyperquader-Klassifikator. Testen Sie Ihr Programm.