Subspace Clustering für die Analyse von CGH Daten

Emmanuel Alexander Müller
emmanuel.mueller@rwth-aachen.de

Lehrstuhl für Datenmanagement und Exploration
Prof. Dr. Thomas Seidl
RWTH Aachen

Abstract: Durch Problemstellungen bei der Anwendung von traditionellen Clustering-Algorithmen auf hochdimensionalen Daten motiviert, wurde im Rahmen meiner Diplomarbeit ein neues algorithmisches Konzept zum effizienten Subspace Clustering entwickelt. Eine mögliche Anwendung dieses Konzeptes stellt die Analyse von CGH Daten dar. Durch Subspace Clustering ist es möglich, Gruppen von Patienten zu identifizieren, deren Genom charakteristische Veränderungen in bestimmten Abschnitten der DNA aufweist. Durch Analyse der so identifizierten Gruppen können Hinweise auf Zusammenhänge zwischen den charakteristischen Veränderungen und den Erkrankungen der Patienten erkannt werden.

Zuordnung: Data Mining, Clustering, hochdimensionale Daten, Comparative Genomic Hybridization

Subspace Clustering

Heutzutage werden enorme Datenmengen zur Analyse von verschiedensten Sachverhalten herangezogen, sei es zur Analyse von Kundenverhalten im Einzelhandel oder Hochdurchsatzexperimenten in der Biologie um nur zwei Beispiele zu nennen. Dabei entstehen neue Problemstellungen, welche nur durch den Einsatz von effizienten Data Mining Verfahren gelöst werden können. Clustering ist eines der hierzu verwendeten Verfahren, wie z.B. in DBSCAN [EKSX96] umgesetzt. Clustering ist die automatische Gruppierung von Objekten. Objekte innerhalb einer Gruppe sollen möglichst ähnlich sein, während Objekte verschiedener Gruppen möglichst unähnlich sein sollen. Jedoch gelingt es bei manchen Datensätzen, welche sich durch einen hochdimensionalen Datenraum auszeichnen, nicht mehr mit den bisher gebräuchlichen Clustering Algorithmen gute Resultate zu erzielen. Dies liegt in der hohen Dimensionalität (Anzahl der Attribute) der verwendeten Objektbeschreibung begründet. Für traditionelle Clustering Verfahren erscheinen aufgrund der hohen Dimensionalität die gegeben Daten als gleichverteilt, wodurch keine Cluster zu erkennen sind. Irrelevante Dimensionen, in denen die Objekte einer Gruppe stark streuen, sind dafür verantwortlich.

Eine Lösung für das hier beschriebene Problem scheint einfach. Die irrelevanten Dimensionen sind auszublenden. Als Subspace Cluster (O, S) wird allgemein eine Objektmenge O zusammen mit einer Auswahl von Dimensionen S bezeichnet. Der durch die Dimensionen in S gebildete Teilraum gibt dabei an, welches die relevanten Dimensionen für den durch die Objekte in O gebildeten Cluster sind. Es lassen sich somit Cluster in unterschiedlichen Projektionen des Datenraumes finden, wie erstmals in CLIQUE [AGGR98] vorgestellt. Diese Teilräume und die darin enthaltenen Cluster zu identifizieren, ist die Aufgabenstellung für das Subspace Clustering.

Die naive Reduktion des Problems auf das traditionelle Clustering würde alle möglichen Teilräume mit einem bestehenden Clustering Verfahren untersuchen. Wegen der exponentiellen Anzahl an

möglichen Teilräumen ist dies jedoch nicht effizient möglich. Im Rahmen meiner Diplomarbeit wurde ein effizienter Subspace Clustering Algorithmus entwickelt, welcher auf erfolgreich angewendeten dichte-basierten Ansätzen aufbaut und diese durch eine statistisch fundierte Modellierung erweitert. Durch den Algorithmus wird die Vollständigkeit des Resultates gewährleistet, im Gegensatz zu bestehenden Verfahren, welche nur durch ein approximatives Vorgehen eine effiziente Berechnung erreichen. Durch Verwendung einer neuen Datenstruktur wird im eigenen Algorithmus sowohl die Effizienz als auch die Vollständigkeit bei der Berechnung erzielt.

Subspace Clustering von CGH Daten

Subspace Clustering wurde im Rahmen der Diplomarbeit für die Analyse von CGH (comparative genomic hybridization) Daten verwendet. Der untersuchte Datensatz besteht aus einer Menge von CGH-Untersuchungen von verschieden Patienten, die an Krebs erkrankt sind. Die Daten stammen von der Website Progenetix.net [BC01], welche solche Untersuchungen in einer Datenbank verwaltet und zur Forschung frei zur Verfügung stellt. Bei einer CGH-Untersuchung wird das Genom auf veränderte Genomabschnitte untersucht. Es können dabei Vervielfältigungen oder Deletionen von Abschnitten auf der DNA identifiziert werden. Eine solche Veränderung weist auf das evtl. Vorkommen von Onkogenen (Zugewinn) bzw, Tumorsuppressorgenen (Verlust) in der betroffenen Region hin [LMC$^+$06]. Es soll mit solchen Untersuchungen der Zusammenhang zwischen Mutationen in gewissen Regionen der DNA und verschiedenen Krebserkrankungen untersucht werden. Zu jedem Patienten liegen in dem verwendeten Datensatz für 862 Abschnitte des Genoms Messwerte zur Veränderung der DNA vor. Zusätzlich liegt zu jedem Patienten eine Einteilung in eine Klasse von Krebserkrankungen vor.

Ziel ist es, eine Gruppierung der Patienten anhand der gegebenen CGH-Untersuchungen vorzunehmen. Eine Gruppe soll dabei eine für sie charakteristische Veränderung bestimmter Abschnitte der DNA aufweisen. Die für diese Gruppierungen durch das Subspace Clustering gewählten Dimensionen (Abschnitte der DNA) werden dann, zusammen mit den in diesen Gruppen gefunden Krebserkrankungen, von Experten analysiert. Es kann dadurch ein Hinweis auf einen Zusammenhang zwischen Mutationen in bestimmten Regionen des Genoms und den dabei auftretenden Krebserkrankungen erkannt werden.

Modellierung

Die Veränderungen in den 862 Abschnitten werden aggregiert für die 48 Chromosomenarme betrachtet. Dabei werden Vervielfältigungen, Deletionen und keine Veränderungen oder keine eindeutigen Veränderungen auf einem Chromosomenarm durch die diskreten Attributwerten 1,-1 und 0 modelliert. Durch Subspace Clustering soll eine charakteristische Veränderung (1,-1) auf bestimmten Chromosomenarmen identifiziert werden. Man ist dabei jedoch nicht an keinen Veränderungen oder an nicht eindeutigen Veränderungen, wie sie der Wert 0 repräsentiert, interessiert. Relevante Informationen stellen in diesem Anwendungsfall also nur die Werte 1 und -1 dar, der Wert 0 ist als irrelevant zu betrachten. Als irrelevant wird beim Subspace Clustering eine Dimension mit starker Streuung der Daten angesehen. Man transformiert somit die Werte in drei getrennt voneinander vorliegende Bereiche in einem kontinuierlichen Wertebereich. Der Wert 0 wird dabei in einen Bereich, in welchem die Werte durch einen Zufallsgenerator gleichverteilt vorliegen, transformiert. Abbildung 1 verdeutlicht diese Transformation zweier Objekte mit nur zwei Dimensionen. Es sind pro Dimension die drei Bereiche für die Werte -1,0,1 hervorgehoben, in die ein Objekt transformiert wird. In dem schraffierten Bereich können somit alle Objekte mit Wert 1 in der ersten Dimension als Subspace Cluster erkannt werden.

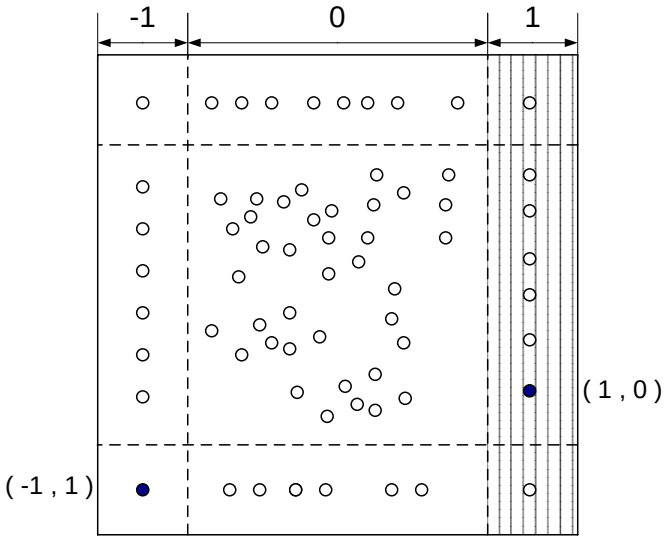

Abbildung 1: Transformation zweier 2D CGH-Objekte zur Identifikation als Subspace Cluster

Diskussion der Ergebnisse

Es wurden Subspace Cluster mit mindestens 20 Objekten (Patienten) im Datensatz gesucht. Dabei wurden Cluster in bis zu 7-dimensionalen Teilräumen (Auwahl von relevanten Chromosomenarmen) gefunden, welche 1407 der 1823 Patienten beinhalten. Die identifizierten Subspace Cluster stellen eine Gruppierung der Krankheitsbilder dar, wobei in den meisten Fällen innerhalb einer Gruppe eine Krankheit den Cluster dominiert. Eine Auswahl von identifizierten Subspace Clustern wurde manuell auf bekannte Zusammenhänge zwischen Mutationen und Erkrankungen untersucht. Dabei konnten schon bekannte Zusammenhänge durch das Clustering bestätigt werden. Darüber hinaus könnte, durch weitere Analysen der gefundenen Subspace Cluster durch Mediziner auf noch unbekannte Zusammenhänge zwischen Mutationen auf bestimmten Chromosomen und gewissen Krebserkrankungen geschlossen werden.

Danksagung
Mein besonderer Dank gilt Michael Baudis für die Bereitstellung und Aufarbeitung der in den Untersuchungen verwendeten CGH Daten.

Literatur

[AGGR98] R. Agrawal, J. Gehrke, D. Gunopulos und P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD*, Seiten 94–105, 1998.

[BC01] M. Baudis und M. Cleary. Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, 17 (12):1228–1229, 2001.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander und X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *KDD*, Seiten 226–231, 1996.

[LMC+06] J. Liu, J. Mohammed, J. Carter, S. Ranka, T. Kahveci und M. Baudis. Distance-based clustering of CGH data. *Bioinformatics*, 22 (16):1971–1978, 2006.