Adaptive Multilevel Cluster Analysis by Self-Organizing Box Maps

Dissertation von

Tobias Galliat

eingereicht am
Fachbereich Mathematik und Informatik
der Freien Universität Berlin

im März 2002

Betreuer:

Prof. Dr. Dr. h.c. Peter Deuflhard
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustr. 7
14195 Berlin

Gutachter:

Prof. Dr. Dr. h.c. Peter Deuflhard
Prof. Dr. Peter Rentrop

Datum der Disputation:

10.07.2002

To my parents

Contents

Zusammenfassung

Als Cluster Analyse bezeichnet man den Prozess der Suche und Beschreibung von Gruppen (Clustern) von Objekten, so daß die Objekte innerhalb eines Clusters bezüglich eines gegebenen Maßes maximal homogen sind. Die Homogenität der Objekte hängt dabei direkt oder indirekt von den Ausprägungen ab, die sie für eine Anzahl festgelegter Attribute besitzen. Die Suche nach Clustern läßt sich somit als Optimierungsproblem auffassen, wobei die Anzahl der Cluster vorher bekannt sein muß. Wenn die Anzahl der Objekte und der Attribute groß ist, spricht man von komplexen, hoch-dimensionalen Cluster Problemen. In diesem Fall ist eine direkte Optimierung zu aufwendig, und man benötigt entweder heuristische Optimierungsverfahren oder Methoden zur Reduktion der Komplexität. In der Vergangenheit wurden in der Forschung fast ausschließlich Verfahren für geometrisch basierte Clusterprobleme entwickelt. Bei diesen Problemen lassen sich die Objekte als Punkte in einem von den Attributen aufgespannten metrischen Raum modellieren; das verwendete Homogenitätsmaß basiert auf der geometrischen Distanz der den Objekten zugeordneten Punkte. Insbesondere zur Bestimmung sogenannter metastabiler Cluster sind solche Verfahren aber offensichtlich nicht geeignet, da metastabile Cluster, die z.B. in der Konformationsanalyse von Biomolekülen von zentraler Bedeutung sind, nicht auf einer geometrischen, sondern einer dynamischen Ähnlichkeit beruhen.

In der vorliegenden Arbeit wird ein allgemeines Clustermodell vorgeschlagen, das zur Modellierung geometrischer, wie auch dynamischer Clusterprobleme geeignet ist. Es wird eine Methode zur Komplexitätsreduktion von Clusterproblemen vorgestellt, die auf einer zuvor generierten Komprimierung der Objekte innerhalb des Datenraumes basiert. Dabei wird bewiesen, daß eine solche Reduktion die Clusterstruktur nicht zerstört, wenn die Komprimierung fein genug ist. Mittels selbstorganisierter neuronaler Netze lassen sich geeignete Komprimierungen berechnen. Um eine signifikante Komplexitätsreduktion ohne Zerstörung der Clusterstruktur zu erzielen, werden die genannten Methoden in ein mehrstufiges Verfahren eingebettet. Da neben der Identifizierung der Cluster auch deren effiziente Beschreibung notwendig ist, wird ferner eine spezielle Art der Komprimierung vorgestellt, der eine Boxdiskretisierung des Datenraumes zugrunde liegt.

Diese ermöglicht die einfache Generierung von regelbasierten Clusterbeschreibungen. Für einen speziellen Typ von Homogenitätsfunktionen, die eine stochastische Eigenschaft besitzen, wird das mehrstufige Clusterverfahren um eine Perroncluster Analyse erweitert. Dadurch wird die Anzahl der Cluster, im Gegensatz zu herkömmlichen Verfahren, nicht mehr als Eingabeparameter benötigt. Mit dem entwickelten Clusterverfahren kann erstmalig eine computergestützte Konformationsanalyse großer, für die Praxis relevanter Biomoleküle durchgeführt werden. Am Beispiel des *HIV Protease Inhibitors VX-478* wird dies detailliert beschrieben.

Lebenslauf

Persönliche Daten

Name:	Galliat
Vorname:	Tobias
geboren am:	17.11.1972 in Köln
Familienstand:	ledig
Konfession:	röm.-kath.
Staatsangehörigkeit:	deutsch

Ausbildung

1979 - 1992	Schulbesuch (Abitur: 06/1992)
10/1992 - 09/1999	Informatikstudium mit Nebenfach BWL, FernUniversität Hagen (Vordiplom: 08/1995, Diplom: 08/1999)
10/1993 - 09/1998	Mathematikstudium, Universität zu Köln (Vordipolm: 10/1995, Diplom: 07/1998)
10/1996 - 12/1996	Auslandstrimester an der Universität Paris-Orsay, Frankreich, im Rahmen eines Erasmus-Stipendiums

Beruflicher Werdegang

07/1992 - 09/1993	Zivildienst, “Referat für interreligiösen Dialog” im Erzbistum Köln
07/1995 - 09/1996	Studentische Hilfskraft, Risk-Consulting, Prof. Dr. Weyer, Köln
01/1997 - 12/1998	Studentischer Mitarbeiter, Risk-Consulting, Prof. Dr. Weyer, Köln
01/1999 - 09/1999	Wissenschaftlicher Mitarbeiter, Risk-Consulting, Prof. Dr. Weyer, Köln
04/1999 - 09/1999	Übungsbetreuung und Vorlesungsvertretung zum Thema “Neuronale Netze” an der Universität zu Köln (zusammen mit Herrn Prof. Dr. Weyer)
seit 10/1999	Wissenschaftlicher Angestellter, Konrad-Zuse-Zentrum für Informationstechnik Berlin