

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 Deuffhard
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustr. 7
14195 Berlin

Gutachter:

Prof. Dr. Dr. h.c. Peter Deuffhard
Prof. Dr. Peter Rentrop

Datum der Disputation:

10.07.2002

To my parents

Contents

Introduction	3
1 Cluster Analysis in High-Dimensional Data	7
1.1 Modeling	8
1.1.1 Geometric cluster problems	9
1.1.2 Dynamic cluster problems	12
1.2 Problem reduction via representative clustering	13
1.3 Efficient cluster description	16
1.4 How many clusters?	21
2 Decomposition	23
2.1 General Definition	23
2.2 Approximate box decomposition	25
2.3 Decomposition based representative clustering	27
2.4 Efficient cluster description via approximate box decomposition	34
2.4.1 Computation of membership rules	34
2.4.2 Discriminating attributes	36
3 Adaptive Decomposition by Self-Organized Neural Networks	41
3.1 Self-Organizing Maps (SOM)	42
3.2 Self-Organizing Box Maps (SOBM)	44
3.3 Comparison SOM - SOBM	53
3.4 Computational complexity	56
3.5 Practical extensions	57
3.5.1 Pruning	58
3.5.2 Early stopping	58
4 Multilevel Representative Clustering	59
4.1 General approach	59
4.2 Adaptive decomposition refinement	60
4.3 Approach based on Perron Cluster analysis	61

4.3.1	Theoretical background	62
4.3.2	Stochastic homogeneity functions	64
5	Applications	73
5.1	Conformational analysis of biomolecules	73
5.1.1	Introduction	73
5.1.2	Adaptation of SOM and SOBM to cyclic data	76
5.1.3	Numerical results: HIV protease inhibitor	79
5.1.4	Prospect: Virtual screening	86
5.2	Cluster analysis of insurance customers	87
5.2.1	Modeling	87
5.2.2	Numerical results: Whiplash Injury Patients	87
	Conclusion	91
	Appendix	93
	Symbols	95
	Bibliography	97
	Zusammenfassung	103
	Lebenslauf	105

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 Peroncluster 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
(Vordiplom: 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 Vorlesungsververtretung 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

