Diss. ETH No. 17822

Infinite-Dimensional Exponential Families in Cluster Analysis of Structured Data

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences

presented by
Peter Orbanz
Dipl.-Inf. (University of Bonn)
born 4. September 1975
citizen of Germany

accepted on the recommendation of
Joachim M. Buhmann, examiner
Zoubin Ghahramani, co-examiner
Sara van de Geer, co-examiner

2008

Abstract

The principal focus of the present thesis is the cluster analysis of structured data, in particular with spatial and temporal coupling structure, and with ordinal structure of individual observations. The thesis studies grouping models and algorithms which (i) take into account structure such as spatial and temporal smoothness and (ii) estimate or select the model order, or number of clusters, from the data. Applications include segmentation of image and remote sensing data, video sequences, and cluster analysis of rank data.

For the cluster analysis of ordinal data, we introduce a mixture model suitable for the simultaneous representation of partial rankings of different lengths. Unlike grouping models for ordinal data previously available in the literature, the model permits the analysis of heterogeneous data sets. For segment analysis of noisy image data, we show that nonparametric Bayesian mixture models can be combined with Markov random fields. The resulting class of models simultaneously estimates the model order and a segmentation of the image, under a smoothness constraint on the segment solution. Video sequences exhibit a temporal dependence structure, similar to spatial coupling in still images. A sequence model of conditional Dirichlet processes is proposed, based on the conjugate nature of the Dirichlet process, which estimates cluster structure evolving over time. By conditioning on a previous estimate, such methods perform *model order adaptation* rather than model order selection. Based on the observation that the Dirichlet process inherits a number of key properties from its finite-dimensional marginals, we study the general construction of infinite-dimensional ("nonparametric") Bayesian models. It is shown that infinite-dimensional conjugate models are generated as projective limits of finite-dimensional conjugate models, in particular of those with exponential family components. An extension theorem for conditional measures and general construction criteria are given, and sufficiency and conjugacy properties of finite-dimensional Bayesian models are shown to be preserved under extension to the infinite-dimensional case.

Zusammenfassung

Schwerpunkt der vorliegenden Arbeit ist die Gruppierungsanalyse (Cluster-Analyse) strukturierter Daten, insbesondere von Daten mit räumlicher und zeitlicher Kopplungsstruktur, sowie mit ordinaler Struktur einzelner Messungen. Die Dissertation studiert Gruppierungsmethoden welche (i) Strukturen wie räumliche und zeitliche Glattheit in die Analyse miteinbeziehen, und (ii) die *Modellordnung*, d.h. die Anzahl der Gruppen, aus den vorliegenden Daten schätzen. Anwendungen umfassen die Segmentierung von Bildern und Fernerkundungsdaten, von Videosequenzen, und die Gruppierungsanalyse ordinaler Präferenzdaten.

Zur Gruppierungsanalyse von Ordinaldaten wird ein Mixturmodell vorgestellt, welches die simultane Repräsentation partieller ordinaler Messungen verschiedener Länge ermöglicht, und damit die ganzheitliche Analyse heterogener Datensätze, welche mit zuvor in der Literatur verfügbaren Ansätzen nicht möglich ist. Zur Segmentierung verrauschter Bilder wird gezeigt, dass nichtparametrische Bayessche Mixturmodelle mit Markovschen Zufallsfeldern kombinierbar sind. Die resultierende Modellklasse erlaubt die simultane Schätzung der Modellordnung und der Segmente unter einer Glattheitsbedingung an die Segmentierungslösung. In Videosequenzen existiert, ähnlich der räumlichen Kopplungsstruktur in Einzelbildern, eine zeitliche Abhängigkeit. Es wird gezeigt, wie Dirichlet-Prozesse unter Ausnutzung ihrer konjugierten Eigenschaften zur Schätzung von über die Zeit evolvierender Gruppenstruktur eingesetzt werden können. Durch Konditionierung auf die jeweils vorangehende Lösung führen solche Verfahren eine *Adaption* (anstatt einer Selektion) der Modellordnung durch. Wir untersuchen die allgemeine Konstruktion unendlich-dimensionaler ("nichtparametrischer") Bayesscher Modelle, ausgehend von der Beobachtung, dass die Marginalverteilungen des Dirichlet-Prozesses Modelle in der Exponentialfamilie darstellen, und eine Reihe wichtiger Eigenschaften an den Prozess vererben. Es wird gezeigt, dass nichparametrische konjugierte Modelle als projektive Limiten solcher endlich-dimensionaler Modelle entstehen, welche ebenfalls konjugiert sind, und damit insbesondere von Modellen in der Exponentialfamilie. Wir geben einen Fortsetzungssatz für bedingte Verteilungen und allgemeine Konstruktionsbedingungen an, und zeigen, inwiefern Suffizienz- und Konjugiertheits-Eigenschaften endlich-dimensionaler Bayesscher Modelle auf den unendlich-dimensionalen Fall übertragbar sind.