Lecture Notes in Computer Science 3940

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Craig Saunders Marko Grobelnik
Steve Gunn John Shawe-Taylor (Eds.)

Subspace, Latent Structure and Feature Selection

Statistical and Optimization
Perspectives Workshop, SLSFS 2005
Bohinj, Slovenia, February 23-25, 2005
Revised Selected Papers

Volume Editors

Craig Saunders
Steve Gunn
John Shawe-Taylor
University of Southampton
School of Electronics and Computer Science
ISIS Group
Southampton, SO17 1BJ, UK
E-mail: {cjs,srg,jst}@ecs.soton.ac.uk

Marko Grobelnik
J. Stefan Institute
Department of Intelligent Systems
Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Marko.Grobelnik@ijs.si

Library of Congress Control Number: 2006925463

CR Subject Classification (1998): F.1, F.2, I.2.6-8, I.2.3, I.4-6, F.4.1, H.3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-34137-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-34137-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11752790 06/3142 5 4 3 2 1 0

Preface

The inspiration for this volume was a workshop held under the auspices of the PASCAL Network of Excellence. Details of the event and more information about the Network can be found under the heading 'Workshop Organization.' The aim of this preface is to provide an overview of the contributions to this volume, placing this research in its wider context.

The aim of the workshop was to bring together researchers working on subspace and latent variable techniques in different research communities in order to create bridges and enable cross-fertilization of ideas. For this reason we deliberately sought invited refereed contributions that would survey a broader field of research giving a common notation and entry point into the individual papers.

The five invited contributions are (in alphabetical order of first author) Avrim Blum on *Random Projection, Margins, Kernels and Feature Selection*, Wray Buntine and Aleks Jakulin on *Discrete Principal Components Analysis*, Dunja Mladenić on *Dimensionality Reduction by Feature Selection in Machine Learning*, Roman Rosipal and Nicole Krämer on *Overview and Recent Advances in Partial Least Squares*, and Mike Titterington on *Some Aspects of Latent Structure Analysis*.

Blum considers subspace selection by random projection. The theoretical analysis of this approach provides an important bound on the generalization of large margin algorithms, but it can also be implemented in kernel defined feature spaces through a two-stage process. The paper provides a survey of a number of clean and important theoretical results. Buntine and Jakulin consider method of determining latent structure based on probabilistic generative models of the data. Their paper gives an introduction to these advanced and effective methods presented from within the machine learning community. Titterington's contribution is a closely related paper but comes from the statistics tradition providing a general framework within which discrete and continuous combinations of latent and observed variables can be placed. Mladenić considers the restricted class of axis parallel subspaces that correspond to feature selection. There is a long tradition of this approach within machine learning and the paper provides an overview of a range of techniques for selecting features, discussing their weaknesses and carefully evaluating their performance. Rosipal and Krämer give a detailed introduction to partial least squares, an important method of subspace selection developed within the chemometrics research community. It can be thought of as an adaptation of principal components analysis where the projection directions have been chosen to be well-suited for solving a set of regression tasks. The authors discuss the kernelization of the technique together with other more recent results.

The contributed papers cover a range of application areas and technical approaches. Agakov and Barber develop a probabilistic modelling technique with a

novel twist of using encoding models rather than generative ones; Monay et al. again consider computer vision using a probabilistic modelling approach; Navot et al. analyze a simple two Gaussian example to show that feature selection can make significant differences in performance and that techniques such as support vector machines are not able to avoid the difficulties of non-informative features in this case; Bouveyron et al. consider a computer vision application using a probabilistic modelling approach; Gruber and Weiss continue the computer vision theme but introduce prior knowledge to enhance the ability to factorize image data to perform 3D reconstruction; Savu-Krohn and Auer use a clustering approach to reduce feature dimensions for image data; Rogers and Gunn consider random forests as an approach to feature selection; Maurer gives frequentist style generalization bounds on PCA-like subspace methods; and finally Reunanen discusses the biases of using cross-validation to do feature selection and outlines some techniques to prevent the introduction of such a bias.

We commend the volume to you as a broad introduction to many of the key approaches that have been developed for subspace identification and feature selection. At the same time the contributed talks give insightful examples of applications of the techniques and highlight recent developments in this rapidly expanding research area. We hope that the volume will help bridge the gaps between different disciplines and hence enable creative collaborations that will bring benefit to all involved.

February 2006

Marko Grobelnik
Steve Gunn
Craig Saunders
John Shawe-Taylor

Workshop Organization

Many of the papers in this proceedings volume were presented at the PASCAL Workshop entitled *Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimization Perspectives* which took place in Bohinj, Slovenia during February, 23–25 2005.

The workshop was part of a Thematic Programme Linking Learning and Statistics with Optimization that ran over the first half of 2005. The PASCAL Network is a European Network of Excellence funded by the European Union under the IST programme. It currently has around 300 researchers at 55 institutions. Its center of gravity is machine learning, but it aims to build links with both optimization and statistics as well as with a range of application areas. It sponsors and co-sponsors a wide range of workshops either organized independently or co-located with international conferences. More information can be found on the website http://www.pascal-network.org.

The Bohinj workshop was hosted by the Institute Josef Stefan, which provided all of the local organization. We are indebted to them for all of the hard work that they put into making the event such a success, although even they could not have planned the magical winter scene that awaited us on our arrival. Particular thanks are due to Tina Anžič, who handled the reservations and hotel bookings as well as many of the travel arrangements.

This work was supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

Organizing Committee

Marko Grobelnik	Jožef Stefan Institute, Ljubljana, Slovenia
Steve Gunn	ISIS Group, University of Southampton, UK
Craig Saunders	ISIS Group, University of Southampton, UK
John Shawe-Taylor	ISIS Group, University of Southampton, UK

Table of Contents

Invited Contributions

Contributed Papers