Kernel Methods for Pattern Analysis

Pattern Analysis is the process of finding general relations in a set of data, and forms the core of many disciplines, from neural networks to so-called syntactical pattern recognition, from statistical pattern recognition to machine learning and data mining. Applications of pattern analysis range from bioinformatics to document retrieval.

The kernel methodology described here provides a powerful and unified framework for all of these disciplines, motivating algorithms that can act on general types of data (e.g. strings, vectors, text, etc.) and look for general types of relations (e.g. rankings, classifications, regressions, clusters, etc.). This book fulfils two major roles. Firstly it provides practitioners with a large toolkit of algorithms, kernels and solutions ready to be implemented, many given as Matlab code suitable for many pattern analysis tasks in fields such as bioinformatics, text analysis, and image analysis. Secondly it furnishes students and researchers with an easy introduction to the rapidly expanding field of kernel-based pattern analysis, demonstrating with examples how to handcraft an algorithm or a kernel for a new specific application, while covering the required conceptual and mathematical tools necessary to do so.

The book is in three parts. The first provides the conceptual foundations of the field, both by giving an extended introductory example and by covering the main theoretical underpinnings of the approach. The second part contains a number of kernel-based algorithms, from the simplest to sophisticated systems such as kernel partial least squares, canonical correlation analysis, support vector machines, principal components analysis, etc. The final part describes a number of kernel functions, from basic examples to advanced recursive kernels, kernels derived from generative models such as HMMs and string matching kernels based on dynamic programming, as well as special kernels designed to handle text documents.

All those involved in pattern recognition, machine learning, neural networks and their applications, from computational biology to text analysis will welcome this account.

Kernel Methods for Pattern Analysis

John Shawe-Taylor
University of Southampton

Nello Cristianini
University of California at Davis

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Monotype Times 10/13pt *System* LaTeX 2$_\varepsilon$ [AUTHOR]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data available

ISBN 0 521 81397 2 hardback

Contents

Code fragments

Preface

The study of patterns in data is as old as science. Consider, for example, the astronomical breakthroughs of Johannes Kepler formulated in his three famous laws of planetary motion. They can be viewed as relations that he detected in a large set of observational data compiled by Tycho Brahe.

Equally the wish to automate the search for patterns is at least as old as computing. The problem has been attacked using methods of statistics, machine learning, data mining and many other branches of science and engineering.

Pattern analysis deals with the problem of (automatically) detecting and characterising relations in data. Most statistical and machine learning methods of pattern analysis assume that the data is in vectorial form and that the relations can be expressed as classification rules, regression functions or cluster structures; these approaches often go under the general heading of 'statistical pattern recognition'. 'Syntactical' or 'structural pattern recognition' represents an alternative approach that aims to detect rules among, for example, strings, often in the form of grammars or equivalent abstractions.

The evolution of automated algorithms for pattern analysis has undergone three revolutions. In the 1960s efficient algorithms for detecting linear relations within sets of vectors were introduced. Their computational and statistical behaviour was also analysed. The Perceptron algorithm introduced in 1957 is one example. The question of how to detect nonlinear relations was posed as a major research goal at that time. Despite this developing algorithms with the same level of efficiency and statistical guarantees has proven an elusive target.

In the mid 1980s the field of pattern analysis underwent a 'nonlinear revolution' with the almost simultaneous introduction of backpropagation multilayer neural networks and efficient decision tree learning algorithms. These

approaches for the first time made it possible to detect nonlinear patterns, albeit with heuristic algorithms and incomplete statistical analysis. The impact of the nonlinear revolution cannot be overemphasised: entire fields such as data mining and bioinformatics were enabled by it. These nonlinear algorithms, however, were based on gradient descent or greedy heuristics and so suffered from local minima. Since their statistical behaviour was not well understood, they also frequently suffered from overfitting.

A third stage in the evolution of pattern analysis algorithms took place in the mid-1990s with the emergence of a new approach to pattern analysis known as kernel-based learning methods that finally enabled researchers to analyse nonlinear relations with the efficiency that had previously been reserved for linear algorithms. Furthermore advances in their statistical analysis made it possible to do so in high-dimensional feature spaces while avoiding the dangers of overfitting. From all points of view, computational, statistical and conceptual, the nonlinear pattern analysis algorithms developed in this third generation are as efficient and as well founded as linear ones. The problems of local minima and overfitting that were typical of neural networks and decision trees have been overcome. At the same time, these methods have been proven very effective on non vectorial data, in this way creating a connection with other branches of pattern analysis.

Kernel-based learning first appeared in the form of support vector machines, a classification algorithm that overcame the computational and statistical difficulties alluded to above. Soon, however, kernel-based algorithms able to solve tasks other than classification were developed, making it increasingly clear that the approach represented a revolution in pattern analysis. Here was a whole new set of tools and techniques motivated by rigorous theoretical analyses and built with guarantees of computational efficiency.

Furthermore, the approach is able to bridge the gaps that existed between the different subdisciplines of pattern recognition. It provides a unified framework to reason about and operate on data of all types be they vectorial, strings, or more complex objects, while enabling the analysis of a wide variety of patterns, including correlations, rankings, clusterings, etc.

This book presents an overview of this new approach. We have attempted to condense into its chapters an intense decade of research generated by a new and thriving research community. Together its researchers have created a class of methods for pattern analysis, which has become an important part of the practitioner's toolkit.

The algorithms presented in this book can identify a wide variety of relations, ranging from the traditional tasks of classification and regression, through more specialised problems such as ranking and clustering, to ad-

vanced techniques including principal components analysis and canonical correlation analysis. Furthermore, each of the pattern analysis tasks can be applied in conjunction with each of the bank of kernels developed in the final part of the book. This means that the analysis can be applied to a wide variety of data, ranging from standard vectorial types through more complex objects such as images and text documents, to advanced datatypes associated with biosequences, graphs and grammars.

Kernel-based analysis is a powerful new tool for mathematicians, scientists and engineers. It provides a surprisingly rich way to interpolate between pattern analysis, signal processing, syntactical pattern recognition and pattern recognition methods from splines to neural networks. In short, it provides a new viewpoint whose full potential we are still far from understanding.

The authors have played their part in the development of kernel-based learning algorithms, providing a number of contributions to the theory, implementation, application and popularisation of the methodology. Their book, *An Introduction to Support Vector Machines*, has been used as a textbook in a number of universities, as well as a research reference book. The authors also assisted in the organisation of a European Commission funded Working Group in 'Neural and Computational Learning (NeuroCOLT)' that played an important role in defining the new research agenda as well as in the project 'Kernel Methods for Images and Text (KerMIT)' that has seen its application in the domain of document analysis.

The authors would like to thank the many people who have contributed to this book through discussion, suggestions and in many cases highly detailed and enlightening feedback. Particularly thanks are owing to Gert Lanckriet, Michinari Momma, Kristin Bennett, Tijl DeBie, Roman Rosipal, Christina Leslie, Craig Saunders, Bernhard Schölkopf, Nicolò Cesa-Bianchi, Peter Bartlett, Colin Campbell, William Noble, Prabir Burman, Jean-Philippe Vert, Michael Jordan, Manju Pai, Andrea Frome, Chris Watkins, Juho Rousu, Thore Graepel, Ralf Herbrich, and David Hardoon. They would also like to thank the European Commission and the UK funding council EPSRC for supporting their research into the development of kernel-based learning methods.

Nello Cristianini is Assistant Professor of Statistics at University of California in Davis. Nello would like to thank UC Berkeley Computer Science Department and Mike Jordan for hosting him during 2001–2002, when Nello was a Visiting Lecturer there. He would also like to thank MIT CBLC and Tommy Poggio for hosting him during the summer of 2002, as well as the Department of Statistics at UC Davis, which has provided him with an ideal environment for this work. Much of the structure of the book is based on

courses taught by Nello at UC Berkeley, at UC Davis and tutorials given in a number of conferences.

John Shawe-Taylor is professor of computing science at the University of Southampton. John would like to thank colleagues in the Computer Science Department of Royal Holloway, University of London, where he was employed during most of the writing of the book.