Lecture Notes in Artificial Intelligence 5400

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Michael Biehl Barbara Hammer
Michel Verleysen Thomas Villmann (Eds.)

Similarity-Based Clustering

Recent Developments and Biomedical Applications

Springer

Volume Editors

Michael Biehl
University Groningen, Mathematics and Computing Science
Intelligent Systems Group, P.O. Box 407, 9700 AK Groningen, The Netherlands
E-mail: m.biehl@rug.nl

Barbara Hammer
Clausthal University of Technology, Department of Computer Science
38679 Clausthal-Zellerfeld, Germany
E-mail: hammer@in.tu-clausthal.de

Michel Verleysen
Université catholique de Louvain, Machine Learning Group, DICE
Place du Levant, 3-B-1348, Louvain-la-Neuve, Belgium
E-mail: michel.verleysen@uclouvain.be

Thomas Villmann
University of Applied Sciences Mittweida
Dep. of Mathematics/Physics/Computer Sciences
Technikumplatz 17, 09648 Mittweida, Germany
E-mail: thomas.villmann@hs-mittweida.de

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.3.3, I.5.3, I.5.4, J.3, F.1.1, I.2.6

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-01804-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-01804-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12654949 06/3180 5 4 3 2 1 0

A physicist once came to Dagstuhl
and thought: 'The castle is quite cool!'
'So, clearly', he stated,
'we should replicate it!
And learn from one instance the right rule.'

Preface

Similarity-based learning methods have a great potential as an intuitive and flexible toolbox for mining, visualization, and inspection of large data sets. They combine simple and human-understandable principles, such as distance-based classification, prototypes, or Hebbian learning, with a large variety of different, problem-adapted design choices, such as a data-optimum topology, similarity measure, or learning mode. In medicine, biology, and medical bioinformatics, more and more data arise from clinical measurements such as EEG or fMRI studies for monitoring brain activity, mass spectrometry data for the detection of proteins, peptides and composites, or microarray profiles for the analysis of gene expressions. Typically, data are high-dimensional, noisy, and very hard to inspect using classic (e.g., symbolic or linear) methods. At the same time, new technologies ranging from the possibility of a very high resolution of spectra to high-throughput screening for microarray data are rapidly developing and carry the promise of an efficient, cheap, and automatic gathering of tons of high-quality data with large information potential. Thus, there is a need for appropriate machine learning methods which help to automatically extract and interpret the relevant parts of this information and which, eventually, help to enable understanding of biological systems, reliable diagnosis of faults, and therapy of diseases such as cancer based on this information. Moreover, these application scenarios pose fundamental and qualitatively new challenges to the learning systems because of the specifics of the data and learning tasks. Since these characteristics are particularly pronounced within the medical domain, but not limited to it and of principled interest, this research topic opens the way toward important new directions of algorithmic design and accompanying theory.

Similarity-based learning models are ideally suited for an application in medical and biological domains or application areas which incorporate related settings because of several crucial aspects including the possibility of human insight into their behavior, a large flexibility of the models, and an excellent generalization ability also for high-dimensional data. Several successful applications ranging from the visualization of microarray profiles up to cancer prediction demonstrate this fact. However, some effort will still be needed to develop reliable, efficient, and user-friendly machine-learning tools and the corresponding theoretical background to really suit the specific needs in medical applications: on the one hand, this application area poses specific requests on interpretability and accuracy of the models and their usability for people without knowledge in machine learning (e.g., physicians), on the other hand, typical problems and data structures in the medical domain provide additional information (e.g., prior knowledge about proteins, the specific form of a spectrum, or the relevance of certain data features), which could be included into the learning methods to shape the output according to the specific situation.

In Spring 2007, 33 scientist from 10 different countries gathered together in Dagstuhl Castle in the south of Germany to discuss important new scientific developments and challenges in the frame of unsupervised clustering, in particular in the context of applications in life science. According to the interdisciplinary topic, researchers came from different disciplines including pattern recognition and machine learning, theoretical computer science, statistical physics, biology, and medicine. Within this highly stimulating and interdisciplinary environment, several topics at the frontiers of research were discussed concerning a theoretical investigation and foundation of prototype-based learning algorithms, the development and extension of models to new challenging directions such as general data structures, and the application for the domain of medicine and biology. While often tackled separately, these three stages were discussed together to judge their mutual influence and to further an integration of different aspects to achieve optimum models, algorithmic design, and theoretical background. This book has emerged as a summary and result of the interdisciplinary meeting, and it gathers together overview articles about recent developments, trends, and applications of similarity-based learning toward biomedical applications and beyond. In three chapters, the three fundamental aspects of a theoretical background, the representation of data and their connection to algorithms, and their particular challenging applications are considered.

More precisely, the first chapter deals with the objectives of clustering and the dynamics of similarity-based learning. Clustering is an inherently ill-posed problem, and optimal solutions depend on the concrete task at hand. Therefore, a variety of different methods exist that are derived on the basis of general mathematical principles or even only of intuitive heuristics. Correspondingly, an exact mathematical investigation of the methods is crucial to judge the performance and reliability of the methods, but at the same time it is often very difficult. Further, even if a concrete objective of clustering is given, the learning dynamics can be highly nontrivial and algorithmic optimization might be necessary to arrive at good solutions of the problem. The first article by Biehl, Caticha, and Riegler gives an overview of the possibility to exactly investigate the learning dynamics of similarity-based learning rules by means of the so-called theory of online learning, which is heavily based on statistical physics. A specific learning rule, neural gas, constitutes the focus of the second contribution by Villmann, Hammer, and Biehl, and possibilities to extend the dynamics to general metrics as well as the connection of topographic coordination to deterministic annealing are discussed. Finally, Fyfe and Barbakh propose an alternative algorithmic solution to get around local minima which consitute a major problem for popular clustering algorithms such as k-means and self-organizing maps. As an alternative, they adapt reinforcement learning algorithms and dynamic programming toward these tasks.

The second chapter addresses another fundamental problem of clustering: the issue of a correct representation of data. In supervised algorithms, data are observed through the interface of a metric, and this metric has to be chosen appropriately to allow efficient and reliable training. Verleysen, Rossi, and Damien consider

the important issue of selecting relevant features if the euclidean metric is used since, otherwise, a failure due to the accumulation of noise can be observed particularly for high-dimensional data sets as often given in biomedical applications. Instead of the euclidean meric, Pearson correlation can be used for non-normalized data with benefitial effects if, for example, the overall shape is more important than the size of the values. This topic is discussed in applications from bioinformatics by Strickert, Schleif, Villmann, and Seiffert. Often, in biological domains, no explicit (differentiable) metric is available at all, rather, data are characterized by pairwise dissimilarities only. Hammer, Hasenfuss, and Rossi discuss extensions of clustering and topographic mapping toward general dissimilarity data by means of the generalized median. Thereby, particular emphasis is laid on the question of how an algorithmic speed-up becomes possible to cope with realistic data sets. Finally, Giannotis and Tino consider fairly general data structures, sequential and graph-structured data, which can be embedded into similarity-based algorithms by means of general probabilistic models that describe the data.

Similarity-based methods find widespread applications in diverse application domains, including in particular biomedical problems, but also geophysics or technical domains. In all areas, however, a number of challenges still have to be faced and a straightforward application of standard algorithms using the euclidean metric is only rarely possible. This observation is demonstrated in the third chapter, which presents discussions about different challenging real-life applications of similarity-based methods. Merenyi, Tademir, and Zhang explain particular problems that occur when very complex data manifolds have to be addressed as occurs, for example, in remote sensing and geoscience applications. Using large-scale data, they demonstrate the problems of extracting true clusters from the data and of evaluating the trustworthiness of the results. Sanchez and Petkov investigate the problem of defining a relevant problem-dependent metric in a biomedical application in image analysis. Ineterstingly, a correct (nontrivial) choice of the metric allows use of comparably simple subsequent processing and it can constitute the core problem in some settings. The final contribution by Rosen-Zvi, Aharni, and Selbig presents state-of-the-art research when addressing challenging large-scale problems in bioinformatics, in this case the prediction of HIV-1 drug resistance. In an impressive way, the question of how to deal with high dimensionality and missing values, among others, is discussed.

Altogether, these presentations give a good overview about important research results in similarity-based learning concerning theory, algorithmic design, and applications, whereby the character of the overview articles with references to correlated research articles makes the contributions particularly suited for a first reading concerning these topics. Since many challenging problems still lie ahead and research will evolve methods beyond this state of the art, there will certainly be a replication of seminars on this topics.

October 2008

Michael Biehl
Barbara Hammer
Michel Verleysen
Thomas Villmann

Table of Contents

Chapter I: Dynamics of Similarity-Based Clustering

Chapter II: Information Representation

Chapter III: Particular Challenges in Applications