"Similarity-based clustering and its application to medicine and biology"

Michael Biehl, RU Groningen, The Netherlands
Barbara Hammer, TU Clausthal, Germany
Michel Verleysen, Université Catholique du Louvain, Belgium
Thomas Villmann, Universität Leipzig, Germany

25.03.07 - 30.03.07

Abstract

The seminar centered around different aspects of similarity-based clustering including theoretical foundations, new algorithms, innovative applications in life science, and future challenges for the field.

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.'

1 Goals of the seminar

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 classical (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 interprete 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.

The seminar centered around developments, understanding, and application of similarity-based clustering in complex domains related to the life sciences. These methods have a great potential as an intuitive and flexible toolbox for mining,

Dagstuhl Seminar Proceedings 07131
Similarity-based Clustering and its Application to Medicine and Biology
http://drops.dagstuhl.de/opus/volltexte/2007/1117

visualization, and inspection of large data sets since they combine simple and human-understandable principles with a large variety of different, problem adapted design choices. The goal of the seminar was to bring together researchers from Computer Science and Biology to explore recent algorithmic developments, discuss theoretical background and problems, and to identify important applications and challenges of the methods.

2 Structure

33 experts from 10 different countries joined the seminar, including a good mixture of established scientists and promising young researchers working in the field. According to the interdisciplinary topic, researchers from Computer Science and related subjects as well as people working in medical departments or biology came together to further the information flow between algorithmic developments and potential applications in this context. Interestingly, a relatively high percentage of the participants works on subjects related to statistical physics, which offers a powerful mathematical foundation for clustering models. During the week, 33 talks were presented which addressed different aspects of clustering and which were grouped into sessions on the following topics:

- Applications in Medicine
- Clustering and Vector Quantization
- Image Processing and Beyond
- High-dimensional Data Processing
- Topographic Models and Non-standard Metrics
- Sparse Representation
- Associations and Dependencies

The talks were supplemented by vivid discussions based on the presented topics and beyond. A dedicated discussion session centered around problems and perspectives in this field to summarize the insights gained during the week and put it into a number of questions/challenges. The Wednesday afternoon session 'Practical Exercise: Sensoric Lab Session' in form of a visit to Trier and subsequent wine tasting gave ample opportunity to further scientific discussions in a very nice environment until late in the evening.

3 Results

A variety of open problems and challenges came up during the week. Before the seminar, the main challenge of similarity-based clustering in medicine and biology was seen as the problem to adapt similarity-based learning for complex, high-dimensional, and possibly non-euclidean data structures as they occur in these domains. During the discussions a much more widespread and subtle picture emerged, identifying the following topics as central issues for clustering:

- **Feature extraction:** feature selection, alternatively the design of a metric or comparison method, seems to be a pivotal issue in clustering as almost everywhere in machine learning. However, the problem seems to be much more pronounced in clustering due to the nature of the problem as an unsupervised learning problem. It is not clear, whether features should be class dependent or general, how to develop dimensionality reduction methods for very high-dimensional data sets which do not simply degenerate, how to incorporate and define invariances of the clusters, etc. A severe problem consists in the fact that it is not clear how to evaluate feature selections since it is not clear how to evaluate clustering.

- **Cluster evaluation:** Clustering is essentially an ill-posed problem and adequate regularization is not clear at all. Often, clustering is therefore evaluated for supervised classification tasks – which give little insight into the 'true' behavior of the clustering unless one can link the classes to real 'clusters', whatever they might be. Further evaluation measures of clustering vary from robustness, usefullness of results (for people from life science), resampling methods, to semisupervised scenarios. However, all (working) algorithms rely on implicit or explicit assumptions on the setting, which are hard to evaluate and compare.

- **Comparison/Benchmarks:** While there exist lots of classification benchmarks, clustering benchmarks are rare and it is very hard to compare clustering results because they rely on different inherent assumptions. Apart from good (challenging but appropriately preprocessed) data sets with clear objectives for benchmarking clustering, standard evaluation procedures and prototcols are missing in this field.

- **Good sampling:** When addressing real problems, sampling is hardly uniform and standard statistical assumptions do generally not hold which makes algorithmic design and evaluation even harder. This problem is particularly pronounced in biology or medicine where data sets are often inbalanced, or they reveal underlying correlations, e.g. due to temporal aspects. Thus, methods to cope with such situations (or to judge whether machine learning is possible in such situations at all) are needed.

Overal, the presentations and dicussions revealed that similarity-based clustering constitutes a highly evolving field which seems particularly suitable for problems in medicine or biology and which still waits with quite a few open problems for researchers, a central problem being a formalization of goals and implicit regularizations of clustering in the context of medicine and biology.