"Similarity-based learning on structures"

Michael Biehl, RU Groningen, The Netherlands
Barbara Hammer, TU Clausthal, Germany
Sepp Hochreiter, University Linz, Austria
Stefan C. Kremer, University of Guelph, Canada
Thomas Villmann, Hochschule Mittweida, Germany

15.02.09 - 20.02.09

Abstract

The seminar centered around different aspects of similarity-based clustering with the special focus on structures. This included theoretical foundations, new algorithms, innovative applications, and future challenges for the field.

For finding the structure in the data set's smothers
many tools are related like sisters and brothers.
We conclude in the sequel:
All methods are equal!
(But some are more equal than others.)

1 Goals of the Seminar

Similarity-based learning methods have a great potential as an intuitive and flexible toolbox for mining, visualization, and inspection of large data sets across several disciplines. While state-of-the-art methods offer efficient solutions for a variety of problems such as the inspection of huge data sets occuring in genomic profiling, satellite remote sensing, medical image processing, etc. a number of important questions requires further research.

The detection, adequate representation, and comparison of structures turned out to be one key issue in virtually all applications. Frequently, learning data contain structural information such as spatial or temporal dependencies, higher order correlations, relational dependencies, or complex causalities. Thus, learning algorithms have to cope with these data structures. In this context, various qualitatively different aspects can be identified: often, data are represented in a specific structured format such as relational databases, XML documents, symbolic sequences, and the like. Similarity based learning has to identify appropriate preprocessing or similarity measures which facilitate further processing. Several problem formulations are ill-posed in the absence of additional structural information e.g. due to a limited

Dagstuhl Seminar Proceedings 09081
Similarity-based learning on structures
http://drops.dagstuhl.de/opus/volltexte/2009/2038

availability of labeled examples for high dimensional data. The dimensionality of microarray data, mass spectra, or hyperspectral images, for example, usually exceeds the number of labelled examples by magnitudes. Structural information can offer effective means for regularization and complexity reduction. More and more learning tasks require additional structural information instead of simple vectorial outputs such as multiple output values, hierarchies, dependencies, or relational information, as required for the inference of biological networks or the analysis of social graphs, for example.

The aim of the seminar was to bring together researchers who develop, investigate, or apply machine learning methods for structure processing to further advance this important field. The focus has been on advanced methods which have a solid theoretical background and display robust and efficient performance in large-scale interdisciplinary applications.

2 Structure

32 experts from 12 countries joined the seminar representing a good mixture of established scientists and young researchers. According to the interdisciplinary topic researches from computer science, mathematics, physics, and related subjects as well as people working in industry came together to discuss and develop new paradigms in the area of structural data processing and learning on structures. During the week 29 talks and short presentations were given which adress different aspects of similarity based learning on structures which could be grouped into clusters on the following topics:

- Structural data processing for biology and medicine
- theoretical aspects of learning for high-dimensional and structered data
- discrete methods for structured data
- stability and quality assessment of data processing in the context of structures
- mathematical aspects of uncertain decisions
- structure-adapted non-standard metrics
- prototype based classification and learning algorithms for stuctures and structured data

The talks were supplemented by vivid discussions based on the presented topics and beyond. Additionally, the talks were complemented by expert software demonstrations which immediately gave a impressive view onto the ability of the presented methodologies. The evening wine and cheese sessions as well as the Wednesday excursion to a local brewery and the manufactory Villeroy&Boch gave ample opportunity to deepen scientific discussions in a relaxed and stimulating atmosphere.

3 Results

A variety of open problems and challenges came up during the workshop week. In particular, the following topics and their interplay were in the focus of several discussions:

- **feature extraction:** Feature selection is one of the main recent topics in classification. Thereby, the task dependent adaptation of predefined data structure models was in the foci of several talks. The methods range from metric adaptation to information theory based selection schemes. The latter follow the naturally inspired paradigm of sparseness while information flow is maximized. The former metric adaptation based approaches optimize the feature set by minimization of classification accuracy. Thereby, classification accuracy has to be defined carefully to cope with the discontinuity of the usual classification error.
- **cluster generation/evaluation:** Cluster generation and evaluation strongly depend on the underlying predetermined similarity/dissimilarity measure applied to the data. The data may be similar according to one measure but may differ heavily with respect to another one. Hence, the choice is crucial for adequate detection of relevant structures and has to be in agreement with the task at hand. The contributions during the seminar presented various approaches taylored according to the needs of different application related problems. Examples are the metric adaptation in quadratic forms for discriminative low-dimensional class representation, development of adaptive kernels or metric adapted multi-dimensional scaling under the specific aspect of high throughput.
- **graph methods for discrete data:** Clustering and classification on graph structures typically require a huge amount of computational costs. Therefore, adaptive methods for approximative solutions are highly desirable. Here the contributions in the seminare were mainly dedicated to the problem of clustering of graphs under the specific restrictions of optimized granularity (in terms of modularity) and the additional requirement of minimization of crossing edges after projection into the plane. The application areas of such problems range from visualization of social networks to dynamics of epidemies.
- **complexity reduction by utilization of structure:** The complexity of data processing of structured data frequently could be reduced if the specific structural information is taken into account. For example, vectorial representation of functions differs from usual data vectors by the spatial dependencies within the vectors. Yet, the utilization of the Euclidean metric disregards this information. During the workshop several approaches for functional metrics were discussed and how they can incorporated into adaptive methods for data processing.

All in all, the presentations and discussion (often until late at night) revealed that similarity based learning on structures constitutes a highly evolving field.

Significant progress has been achived in recent years and was highlighted during the seminar. Although promising results and approaches were developed, many important problems still await satisfactory practical solutions. For example, the functional aspect of data is not sufficiently exploited in many data processing methods. Another challenge is the sparseness of data in high-dimensional data analysis and adequate processing tools.