09081 Abstracts Collection
Similarity-based learning on structures
— Dagstuhl Seminar —

Michael Biehl[1], Barbara Hammer[2], Sepp Hochreiter[3], Stefan C. Kremer[4] and Thomas Villmann[5]

[1] University of Groningen, NL
m.biehl@rug.nl
[2] TU Clausthal, D
[3] University of Linz, A
hochreit@bioinf.jku.at
[4] University of Guelp, CA
skremer@uoguelph.ca
[5] Universität Leipzig, D
thomas.villmann@hs-mittweida.de

Abstract. From 15.02. to 20.02.2009, the Dagstuhl Seminar 09081 "Similarity-based learning on structures " was held in Schloss Dagstuhl – Leibniz Center for Informatics. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Similarity-based clustering and classification, metric adaptation and kernel design, learning on graphs, spatiotemporal data

09081 Summary – Similarity-based learning on structures

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.

Joint work of: Biehl, Michael; Hammer, Barbara; Hochreiter, Sepp; Kremer, Stefan C.; Villmann, Thomas

Keywords: Similarity-based clustering and classification, metric adaptation and kernel design, learning on graphs, spatiotemporal data

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/2038

Clustering and PCA reveal the migration path of modern humans and subtypes of breast cancer

Gyan Bhanot (Rutgers University - Piscataway, US)

Abstract: Mitochondria are organelles in the cytoplasm which play a key role in producing most of the energy needed by the cell. Their origin is a sybmiosis between proto-cyanobacteria and precursors of eukaryotic cells 2 billion years ago. They have a circular chromosome (mtDNA) which is transmitted exclusively by maternal descent and can be used to trace the history of migrating populations. In 1987, a study of some of the mutations on mtDNA for showed that the origins of modern humans can be traced to a single woman - the so called "mitochondrial Eve" who lived in Africa 150,000 years ago. I will describe the background of this discovery and present new results based on PCA and clustering analysis of complete sequencing of thousands of mtDNA which reveals the details of our migration "Out of Africa" approximately 50K-70K years before present.

We also show that similar methods applied to breast cancer microarray data yield clinically relevant subclasses of disease with distinct biology and rates of relapse (metastasis).

See also: 1. G. Alexe, R. Vijaya-Satya, M. Seiler, D. Platt, T. Bhanot, S. Hui, M. Tanaka, A. J. Levine and G. Bhanot, 'PCA and Clustering Reveal Alternate mtDNA Phylogeny of N and M Clades', 2008, J. Mol Evol, 67 (5), 465-487; 4. G. Alexe, G. S. Dalgin, D. Scanfeld, P. Tamayo, J. Mesirov, C. Delisi, L. Harris, N. Bernard, M. Martel, A. Levine, S. Ganesan, G. Bhanot, 'High Expression of Lymphocyte Associated Genes in node-negative HER2+ breast cancer correlates with lower recurrence rates', 2007, Cancer Research, 67: 10669-10676.

Similarity beyond distances

Ulrich Bodenhofer (University of Linz, AT)

There is hardly any statistical learning paradigm that does not make use of some concept of similarity, regardless of whether explicitly or implicitly. This contribution attempts to discuss different concepts of similarity. We will discuss the relationship between distances and gradual similarity along with the correspondence between the triangle inequality and transitivity. To this end, we will highlight the well-established concept of graded equivalence relations which are neither widely known nor widely used in machine learning. We will repeat some known results on relationships of graded equivalence relations and positive semi-definite kernels. Finally, we provide some ideas how to use these results in the classification of biological sequences.

Keywords: Distances, Graded equivalence relation, Kernels, Machine Learning

Structure validation in clustering by stability analysis

Joachim Buhmann (ETH Zürich, CH)

Partitioning of data sets into groups defines an important processing step for compression, prototype extraction or outlier removal. Various criteria of connectedness or proximity have been proposed to group data according to structural similarity but in general it is unclear which method or model to use. In the spirit of information theory we propose a decision process to determine the extractable information from data conditioned on a hypothesis class of structures. Maximizing the amount of information which can be reliably learned from data in the presence of noise selects appropriate models. Empirical evidence for this model selection concept is provided by cluster validation in bioinformatics and in computer security, i.e., the analysis of microarray data and multilabel clustering of Boolean data for role based access control.

Keywords: Clustering, model selection, information theory

Limited Rank LVQ and Applications

Kerstin Bunte (University of Groningen, NL)

We propose an extension of the recently introduced Generalized Matrix Learning Vector Quantization (GMLVQ) algorithm. The original algorithm provides a discriminative distance measure of relevance factors, aided by adaptive square matrices, which can account for correlations between different features and their importance for the classification. We extend the scheme to matrices of limited rank corresponding to low-dimensional representations of the data. This allows to incorporate prior knowledge of the intrinsic dimension and to reduce the number of adaptive parameters efficiently. The case of two- or three-dimensional representations constitutes an efficient visualization method. The identification of a suitable projection is not treated as a pre-processing step but as an integral part of the supervised training. We will also introduce some applications in the field of Content Based Image Retrieval (CBIR) and nonlinear visualization based on the global alignment of local linear representations.

Keywords: Machine Learning, Learning Vector Quantization, Adaptive Distancemeasures, Visualization, Content based Image Retrieval, nonlinear Visualization

"Searching in Equivalence Classes"

Hans Burkhardt (Universität Freiburg, DE)

Searching for identical objects is a rather trivial task. However, it is much more challenging to search in semantic equivalence classes.

In many pattern recognition problems images have to be classified independent of their current position and orientation, which is just a nuisance parameter. Instead of comparing a measured pattern in all possible locations against the prototypes it is much more attractive to extract position-invariant and intrinsic features and to classify the objects in the feature space. Mathematically speaking, patterns form an equivalence class with respect to a geometric coordinate transform describing motion. Invariant transforms are able to map such equivalence classes into one point of an appropriate feature space.

The talk will describe new results for this classical problem and outlines general principles for the extraction of invariant features from images (Haar integrals, Lie-Theory, Normalization techniques). The nonlinear transforms are able to map the object space of image representation into a canonical frame with invariants and geometrical parameters. Beside the mathematical definition the talk will concentrate on characterizing the properties of the nonlinear mappings with respect to completeness and possible ambiguities, disturbance behaviour and computational complexity. We especially investigated Haar integrals for the extraction of invariants with respect to mathematical groups based on monomial and relational kernel functions.

Examples and applications will be given for content-based image retrieval tasks in 2D and 3D (images and voxel data).

Outline:

1. Introduction and Fundamentals
2. Equivalence Classes based on geometric transforms (Euclidean motion, affine mapping)
3. General Principles for the Construction of Invariants
4. Current developments and future perspectives
   - SIMBA - Search Images by Appearance
   - MICHELscope - a search engine for philatelists
   - Search engine for Watermarks in old paper prints
   - I-SEARCH - BMBF-Project on CBIR (patch-based search and relevance feedback)
5. Challenges in Biology
   - Searching and classification of 3D-Structures (Pollen monitor)
   - Searching in Biological databases (Protein search engine)

Keywords: Image retrieval, invariants

Full Paper:
http://lmb.informatik.uni-freiburg.de/papers/index.de.html

FARMS: a probabilistic latent variable model for summarizing Affymetrix array data at probe level

Djork-Arne Clevert (University of Linz, AT)

Motivation:

High-density oligonucleotide microarrays, and in particular Affymetrix GeneChip arrays, are successfully applied in many areas of biomedical research. However, the large number of genes, which have high variation but are irrelevant for the experimental question, leads to many false positives in extracting relevant genes. The high variation which origins from measurement noise can be reduced by identifying the probes of a probe set which vary synchronously across the arrays. The cause for synchronous changes is the gene-specific mRNA concentration in the cell detected by those probes.

This concentration is the hidden factor in our factor analysis model, which automatically models the probe-specific measurement error and is optimized by Bayesian maximum a posteriori estimation.

In contrast to previous methods our new summarization method called "Factor Analysis for Robust Microarray Summarization" (FARMS) supplies model-based signal intensity values.

Results:

We have evaluated FARMS in terms of sensitivity and specificity on all public available spike-in data sets and at the Affycomp competition where it outperformed all competitors. Therefore, FARMS can be very important for biologists and medical researchers, which have to face the problem of false discovery rates. This problem is considerably reduced because genes are filtered out without looking at conditions; therefore higher significance values are obtained in the post-processing.

Keywords: Affymetrix, factor analysis, microarray, summarization, affycomp, probabilistic latent variable model, probe level data

Joint work of: Clevert, Djork-Arne; Hochreiter, Sepp

Full Paper:
http://bioinformatics.oxfordjournals.org/cgi/content/short/22/8/943

See also: @articleSeppHochreiter04152006, author = Hochreiter, Sepp and Clevert, Djork-Arne and Obermayer, Klaus, title = A new summarization method for affymetrix probe level data, journal = Bioinformatics, volume = 22, number = 8, pages = 943-949, doi = 10.1093/bioinformatics/btl033, year = 2006, abstract = Motivation: We propose a new model-based technique for summarizing high-density oligonucleotide array data at probe level for Affymetrix GeneChips. The new summarization method is based on a factor analysis model for which a Bayesian maximum a posteriori method optimizes the model parameters under the assumption of Gaussian measurement noise. Thereafter, the RNA concentration is estimated from the model. In contrast to previous methods our new

method called Factor Analysis for Robust Microarray Summarization (FARMS)' supplies both P-values indicating interesting information and signal intensity values. Results: We compare FARMS on Affymetrix's spike-in and Gene Logic's dilution data to established algorithms like Affymetrix Microarray Suite (MAS) 5.0, Model Based Expression Index (MBEI), Robust Multi-array Average (RMA). Further, we compared FARMS with 43 other methods via the Affycomp II' competition. The experimental results show that FARMS with default parameters outperforms previous methods if both sensitivity and specificity are simultaneously considered by the area under the receiver operating curve (AUC). We measured two quantities through the AUC: correctly detected expression changes versus wrongly detected (fold change) and correctly detected significantly different expressed genes in two sets of arrays versus wrongly detected (P-value). Furthermore FARMS is computationally less expensive then RMA, MAS and MBEI. Availability: The FARMS R package is available from
http://www.bioinf.jku.at/software/farms/farms.html
Contact: hochreit@bioinf.jku.at
Supplementary information:
http://www.bioinf.jku.at/publications/papers/farms/supplementary.ps,
URL = http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/8/943,
eprint = http://bioinformatics.oxfordjournals.org/cgi/reprint/22/8/943.pdf

Kernels and structured output algorithms for prediction of protein metal binding sites

Paolo Frasconi (University of Firenze, IT)

Metalloproteins are involved in many biological process, including apoptosis and aging, and diseases for which medicine is still seeking effective treatment, including Alzheimer. Parkinson, and AIDS. High-throughput technologies can only determine if a protein binds a metal ion but cannot identify the involved residues. The prediction server METALDETECTOR (http://metaldetector.dsi.unifi.it/) can identify the metal bonding state of cysteines and histidines and simultaneously predicts disulfide bridges. A much more challenging task, however, is to identify the complete geometry of binding sites (i.e. which residues are jointly involved in the coordination of each ion) starting from sequence alone. We suggest a formulation based on structured output learning and introduce specialized kernel functions for measuring the similarity between binding geometries in two chains, incorporating both structural and evolutionary information.

Metal binding can be shown to have the algebraic property of a matroid, enabling the application of a novel greedy algorithm for solving the structured output learning problem. The kernels can be extended to the case where three-dimensional information about the protein is available, by enriching the residue similarity with information derived from triads of donor atoms. This setting can be important for example to identify binding sites in known structures (or their

close homologs) that have been experimentally determined in apo (unbound) form.

Keywords: Kernels on structured data, structured output learning, metal binding sites in proteins

Joint work of: Frasconi, Paolo; Passerini, Andrea; Lippi, Marco

Full Paper:
http://www.dsi.unifi.it/∼paolo/publications.html

See also: Frasconi, P. and Passerini, A. (2008). Predicting the Geometry of Metal Binding Sites from Protein Sequence. In 22nd Annual Conference on Neural Information Processing Systems (NIPS'08). Lippi, M., Passerini, A., Punta, M., Rost, B., and Frasconi, P. (2008). MetalDetector: a web server for predicting metal binding sites and disulfide bridges in proteins from sequence. Bioinformatics.

Extending Fuzzy c-Means to handle Similarity Based Data

Tina Geweniger (Universität Leipzig, DE)

Clustering of objects is a main task in machine learning. Thereby one can distinguish between crisp and fuzzy variants. For both strategies advanced methods are developed. Popular approaches favor prototype based adaptive algorithms. Usually , these algorithms belong to the class of vector quantizers, requiring the data objects and the prototypes to be embedded in a metric vector space. Recent designs relax this last condition only demanding the existence of a similarity relation between the data objects given by a data similarity matrix D. Yet, these approaches are only crisp quantizer.

I am going to show an extention of the standard fuzzy c-means algorithm to handle data objects, if only similarities between them are given, i.e. the median variant of FCM.

Keywords: Median fuzzy c-Means, simililarity data

Gene Expression Profiling Using Affymetrix Microarrays

Hinrich Göhlmann (Johnson & Johnson Pharmaceutical - Beerse, BE)

In contrast to most participants of the seminar, I have attended "Similarity-based learning on structures" as a molecular biologist generating the kind of data that is used for developing some of the methods that are presented here. Beyond the biological side of things, I am interested in unsupervised projection methods such as spectral mapping as well as techniques that summarize the multiple measurements per transcript into a single number when looking at the Affymetrix microarray platform for gene expression studies and DNA copy number analysis. I am furthermore interested in learning as much as possible about the whole process - trying to not only understand the biology, but increasingly more the fields of data analysis, statistics, bioinformatics and IT.

Keywords: Affymetrix, Gene expression, Unsupervised projection methods, Gene filtering

Full Paper:
http://goehlmann.info/Research

Bayesian DNA sequence analysis using the same prior

Jens Keilwagen (IPK Gatersleben, DE)

During the last decade, many algorithms for the recognition of short DNA signal sequences utilizing different models and different training principles have been developed and compared with the goal of improving their performance and to deepen our insight into gene regulation at the cellular level. Comparative studies of different approaches help to get a better understanding of the strengths and weaknesses of each approach and do therefore point out possible further improvements of these algorithms. While in most of these studies the choice of the underlying model or training principle has been compared, the influence of choosing different priors has been overlooked in the past. However, neglecting the influence of choosing different priors leads to a comparison of apples and oranges in many cases, and thus to questionable conclusions. Many different models have been proposed for the recognition of short DNA signal sites, including position weight matrix models, weight array matrix models, Bayesian trees, or moral Bayesian networks, but interestingly all of these models are special cases of Markov random fields. We derive a prior for Markov random fields that permits a fair comparison of these models using generative as well as discriminative training principles.

Keywords: Probabilistic models, Markov random fields, DNA sequence analysis, discriminative learning, generative learning

Neural Grammar Networks for Bio-molecular Classification

Stefan C. Kremer (University of Guelph, CA)

Traditional methods for classification in statistics and machine learning often assume that the data to be classified is representable by vectors. While problems in bioinformatics can use encoding schemes to convert sequences and molecules into (feature) vector representations, these techniques by necessity are lossy and require domain specific expertise to identify suitable formulations of features. In this talk, I will present a novel method for encoding molecular structure and presenting it to a classification system. The technique combines previously developed grammatical representations for describing structure with a classification system that is assembles "on-the-fly" on a per molecule basis. This allows even a novice user to exploit the expert knowledge already embedded in the grammatical representations in order to build classifiers that are competitive with existing methods.

Keywords: Neural Grammar Networks, Bio-molecular Classification, QSAR

Full Paper:
http://www.cis.uoguelph.ca

See also: Bibliography

Discovering and exploiting structure for better learning and inference

Erzsebet Merenyi (Rice University, US)

I present two studies.

In the first study, the task is to predict interdependent implicit variables, in this case temperature and grain size, from much higher dimensional spectral measurements of icy planetary surfaces (explicit functional data). We observe an interesting duality of the influences of the two underlying parameters on the structure of the spectra, which counteracts the simultaneous learning of the two parameters, resulting in relatively poor prediction. From the SOM that serves as a middle layer in the supervised ANN, we can get clues about the adverse effect on the learning. Exploiting this knowledge we propose the concept of "conjoined twin machines" which share the same SOM, and two separate "heads" interpret the SOM's knowledge somewhat differently, which allows each to specialize on the prediction of one of the two implicit parameters. This increases the overall accuracy of the prediction (for both parameters) significantly.

The second study looks at a case of misuse of SOMs. While technically simple, it has potentially far reaching consequences. In GENECLUSTER, a software package developed in, and used by the medical community for clustering genes, knowledge of structure in the data, as learned and conveyed by a SOM, is ignored to some extent. This leads to erroneous segmentation, and thus misinterpretation, of groups of similarly expressed genes across time. I re-analyze a data set from which results obtained with GENECLUSTER were published earlier in PNAS. I show cleaner and richer results with proper utilization of the SOM.

Keywords: Parameter prediction, functional data, clustering, Self-Organizing Maps, planetary spectra, gene microarrays

An adaptive model for learning molecular endpoints

Gianluca Pollastri (University College - Dublin, IE)

I will describe a recursive neural network that deals with undirected graphs, and its application to predicting property labels or activity values of small molecules.

The model is entirely general, in that it can process any undirected graph with a finite number of nodes by factorising it into a number of directed graphs with the same skeleton.

The model's only input in the applications I will present is the graph representing the chemical structure of the molecule. In spite of its simplicity, the model outperforms or matches the state of the art in three of the four tasks, and in the fourth is outperformed only by a method resorting to a very problem-specific feature.

Keywords: Recursive neural networks, qsar, qspr, small molecules

Joint work of: Walsh, Ian; Vullo, Alessandro; Pollastri, Gianluca

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2009/2036

Topologically Ordered Graph Clustering

Fabrice Rossi (CNRS - TELECOM-Paristech, FR)

This talk presents an alternative solution to kernel/dissimilarity methods for clustering a graph into communities that are easy to visualize. The main principle of this solution is to modify the most popular quality measure for community detection, Girvan and Newman's modularity, so as to take into account a prior structure, as in the self-organizing map.

We first recall the modularity measure and then introduce the proposed topological variant. The optimization of both measures is a combinatorial problem. We show how to use deterministic annealing to obtain a fast and reliable solution. We show what type of results can be obtained in this framework on real world graphs and outline advantages and drawbacks of the method.

Keywords: Modularity, Graph clustering, Deterministic Annealing, Self-Organizing Map

Joint work of: Rossi, Fabrice; Villa, Nathalie

Structured prediction with kernels: lessons learned and recent progress

Juho Rousu (University of Helsinki, FI)

In my talk I will present some observations and lessons learned in the development of kernel-based structured prediction methods, adn highlight recent progress in an important bioinformatics application of enzyme function prediction.

Keywords: Kernel methods, structured prediction, bioinformatics

Hyperparameter Learning in Robust Soft LVQ

Petra Schneider (University of Groningen, NL)

Learning vector quantization (LVQ) constitutes a powerful and intuitive method for adaptive nearest prototype classification. However, original LVQ has been introduced based on heuristics and numerous modifications exist to achieve better convergence and stability. The Robust Soft LVQ (RSLVQ) algorithm is derived from an explicit cost function and follows the dynamics of a stochastic gradient ascent. The cost function involves a hyperparameter which has strong influence on the performance of the resulting classifier. In this talk, I will present a well-founded strategy to deal with the algorithms' hyperparameter. The new variant of RSLVQ will be demonstrated based on artificial data and a real-life data set.

Evaluation of topography for Evolving Trees

Stefan Simmuteit (Universität Leipzig, DE)

For self-organizing maps exist different measures for the degree of topology preservation. In this talk we focus on the transfer of these concepts to Evolving Trees. In particular the well-known topographic product is carried over to Evolving Trees, because it is robust and approximates the exact mathematical model in a good manner for reasonable computational effort.

The tree-structured neighborhood of Evolving Trees is supposed to conserve the topology of taxonomic data like bacterial mass spectrometry data better than commonly used flat maps. The topographic product is adopted to Evolving Trees and results show a good topology preservation with various parameter constellations, when learning the structure of different bacteria species.

Generalized analysis of data attribute variability

Marc Strickert (IPK Gatersleben, DE)

Many data analysis methods make use of implicit assumptions on the data boundaries. In high-throughput intensity data, such as obtained by cDNA arrays, 2D protein gels, or mass-spectroscopy, the most prominent reference boundary is the theoretically smallest value of zero. For ratio-based analysis of data attribute variability such a boundary has quite an impact on the results. For example, a doubling of temperature from 10 to 20 degrees Celsius relative to zero is just an 1.04 times fluctuation in case of the Kelvin reference point. Data centers of gravity are one way of defining more data-specific reference points. Attribute variability characterized by the mean-centered sums of squares work fine for Euclidean spaces, but not necessarily so for other measures of data similarity.

By the example of Pearson correlation a more general approach to the analysis of attribute variability is presented that allows a less biased assessment of data patterns and attribute structures.

Keywords: Generalized variance, correlation measure

A closer look on codebooks for object class recognition

Alexandra Teynor (Universität Freiburg, DE)

Codebooks are a vital part of many current image retrieval and object class recognition algorithms, especially in so called bag-of-feature settings. Several issues have to be considered when generating such dictionaries. In this talk, we take a deeper look at possible generation processes and show methods how to create semantic instead of purely visual codebooks. The presented techniques show that it is possible to create powerful codebooks with relatively minor effort and that the grouping of visually distinct parts that are semantically equivalent improves the performance.

Keywords: Visual codebook, object class recogniton

Searching for co-expression patterns in three-color cDNA microarray data using a probabilistic model based Hough Transform

Peter Tino (University of Birmingham, GB)

The effects of a drug on the genomic scale can be assessed in a three-color cDNA microarray with the three color intensities represented through the so-called hexaMplot. In our recent study we have shown that the Hough Transform (HT) applied to the hexaMplot can be used to detect groups of co-expressed genes in the normal-disease-drug samples.

However, the standard HT is not well suited for the purpose because: (1) the essayed genes need first to be hard-partitioned into equally and differentially expressed genes, with HT ignoring possible information in the former group; (2) the hexaMplot coordinates are negatively correlated and there is no direct way of expressing this in the standard HT and (3) it is not clear how to quantify the association of co-expressed genes with the line along which they cluster.

We address these deficiencies by formulating a dedicated probabilistic model based HT.

The approach is demostrated by assessing effects of the drug Rg1 on homocysteine-treated human umbilical vein endothetial cells. Compared with our previous study we robustly detect stronger natural groupings of co-expressed genes. Moreover, the gene groups show coherent biological functions with high significance, as detected by the Gene Ontology analysis.

Keywords: Three-color cDNA microarray, Hough Transorm, probabilistic modeling

Joint work of: Tino, Peter; Hongya Zha;, Hong Yan

Estimating Time Delay in Gravitationally Lensed Fluxes

Peter Tino (University of Birmingham, GB)

We study the problem of estimating the time delay between two signals representing delayed, irregularly sampled and noisy versions of the same underlying pattern. We propose a kernel-based technique in the context of an astronomical problem, namely estimating the time delay between two gravitationally lensed signals from a distant quasar.

We test the algorithm on several artificial data sets, and also on real astronomical observations. By carrying out a statistical analysis of the results we present a detailed comparison of our method with the most popular methods for time delay estimation in astrophysics. Our method yields more accurate and more stable time delay estimates. Our methodology can be readily applied to current state-of-the-art optical monitoring data in astronomy, but can also be applied in other disciplines involving similar time series data.

Keywords: Time series, kernel regression, statistical analysis, evolutionary algorithms, mixed representation

Joint work of: Tino, Peter; Cuevas-Tello, Juan C.; Raychaudhury, Somak

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/2037

Topographic graph clustering with kernel and dissimilarity methods

Nathalie Villa-Vialaneix (University of Toulouse, FR)

Finding meaningful communities in social network, or, to use a more classical vocabulary, clustering a graph, is a very important problem in social network analysis. For very large graphs, a fast clustering algorithm provides a coarse graining of the graph and can be considered a preprocessing phase before time consuming algorithms. For small to medium size graphs, communities can be analyzed manually (or at least semi-automatically) by specialists in order to figure out the global organization of the underlying social network.

This talk focuses on the second case, when the goal is to display to an end user a visual representation of a social network. The main idea of our work is to find communities that can be displayed easily on a plane. Our method is based on the self-organizing map paradigm.

We first recall briefly some extension of the self-organizing map to kernel and dissimilarity data. Then we give examples of kernels and dissimilarities that can be constructed to compare the nodes of a graph using its link structure. We show what type of results can be obtained in this framework on real world graphs and outline advantages and drawbacks of the method.

Keywords: Self-organizing map, graph, kernel, social networks

Joint work of: Villa, Nathalie; Rossi, Fabrice

Sobolev Norms for Functional Data Analysis

Thomas Villmann (Hochschule Mittweida, DE)

We propose the utilization of Sobolev norms for functional data analysis. Sobolev metric take the shape of the data by incorporration of the derivatives in to account. We show, how Sobolev norms can be used in functional vector quantization or functional principal component analysis based on the Oja-learning.

Keywords: Functional data analysis, vector quantization, principal component analysis, Sobolev norms

Theoretical study of online and offline LVQ

Aree Witoelar (University of Groningen, NL)

In this talk we describe two approaches for a theoretical analysis of Learning Vector Quantization (LVQ) systems in a controlled environment of Gaussian mixtures in high dimensions and a system of two prototypes.

The first part focuses on online LVQ algorithms using windows for the selection of data. Theory of on-line learning allows for an exact description of the training dynamics, yielding typical learning curves, convergence properties and achievable generalization abilities. We compare the performance of several algorithms, including LVQ 2.1, Learning From Mistakes (LFM) and Robust Soft LVQ.

In the second part we apply the statistical physics analysis of offline learning to cost function based LVQ schemes. The analytic approach becomes exact in the limit of high training temperature. We study two cost function related LVQ algorithms and the influence of an appropriate weight decay. In our findings, LFM achieves poor generalization ability, while LVQ 2.1 displays much better performance with a properly chosen weight decay.

Keywords: LVQ, theory of on-line learning, statistical physics

Incorporation of contextual information into recognition processes

Dietlind Zühlke (Fraunhofer Institut FIT - St. Augustin, DE)

Automatic image analysis methods often do not succeed in biological applications concerning high specificity and generalization. Successful ŞmodelsŤ in this respect are human experts. To consider background knowledge in pertinent situations (as instances of a generalized context) is evidently one important aspect of their success. I want to discuss structures and methods that allow us to incorporate a choice of contextual background information into automatic image analysis. Using the example of automatic pollen recognition I will introduce a simple first approach as a basis for discussion.

Keywords: Image analysis, biological applications, background knowledge

Analysis of Robust Soft Learning Vector Quantization and an application to Facial Expression Recognition

Gert-Jan de Vries (Philips Research Lab. - Eindhoven, NL)

Learning Vector Quantization (LVQ) is a popular method for multiclass classification. Several variants of LVQ have been developed recently, of which Robust Soft Learning Vector Quantization (RSLVQ) is a promising one. Although LVQ methods have an intuitive design with clear updating rules, their dynamics are not yet well understood.

In simulations within a controlled environment RSLVQ performed very close to optimal. This controlled environment enabled us to perform a mathematical analysis as a first step in obtaining a better theoretical understanding of the learning dynamics. In this talk I will discuss the theoretical analysis and its results. Moreover, I will focus on the practical application of RSLVQ to a real world dataset containing extracted features from facial expression data.

Keywords: Learning Vector Quantization, Analysis, Facial Expression Recognition

Joint work of: de Vries, Gert-Jan; Biehl, Michael

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/2035