07131 Abstracts Collection
Similarity-based Clustering and its Application to Medicine and Biology
— Dagstuhl Seminar —

Michael Biehl[1], Barbara Hammer[2], Michel Verleysen[3] and Thomas Villmann[4]

[1] Univ. of Groningen, NL
m.biehl@rug.nl
[2] TU Clausthal, DE
hammer@in.tu-clausthal.de
[3] Univ. of Louvain, BE
verleysen@dice.ucl.ac.be
[4] Univ. Leipzig, DE
thomasvillmann@medicine.uni-leipzig.de

Abstract. From 25.03. to 30.03.2007, the Dagstuhl Seminar 07131 "Similarity-based Clustering and its Application to Medicine and Biology" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. 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, prototype-based classifiers, self-organisation, SOM, learning vector quantization, medical diagnosis, bioinformatics

07131 Summary – Similarity based clustering and its application to medicine and biology

This paper summarizes presentations, discussions, and results of the Dagstuhl seminar.

Keywords: Clustering, bioinformatics, medicine

Joint work of: Biehl, Michael; Hammer, Barbara; Verleysen, Michel; Villmann, Thomas

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2007/1117

Dagstuhl Seminar Proceedings 07131
Similarity-based Clustering and its Application to Medicine and Biology
http://drops.dagstuhl.de/opus/volltexte/2007/1119

Analyzing genome tiling microarrays for the detection of novel expressed genes

Rainer Breitling (University of Groningen, NL)

Motivation: Genomic tiling microarrays offer a unique opportunity for detecting novel genes. A major challenge in interpreting genomic tiling array data is the discrimination between expressed and non-expressed probes, based on very noisy hybridization signals. With genome sequencing and annotation efforts in the past years, we have a reasonably well-annotated genome for many organisms. Based on this available genome annotation, we here consider transcription detection as a supervised classification problem and employ various machine learning methods for determining expressed genome areas.

Methodology: Transcript positions are obtained from the most recent genome annotation and used for classifying probes in a training set. Various published transcript detection methods, as well as Support Vector Machines, and Classification and Regression Trees are used to classify probes, and the results are analyzed to identify parameters that influence performance.

Results: First, we show how information from probe signal intensity, mismatch intensity, melting temperature, hybridization correlation and differential expression between multiple samples can all be used for successful classification. We demonstrate how melting temperature, length of exons, expression level, and probe position within exons affect the performance of transcription detection methods. We then identify the optimal setting of maxgap and minrun parameters for building exon segments. Our approach enables us to objectively compare published methods for expression detection as well as several newly suggested machine learning approaches.

We conclude that genome tiling arrays contain sufficient information to generate reliable de novo whole-genome transcript maps, when analyzed with the appropriate tools.

Keywords: Bioinformatics, Genome tiling arrays, microarrays, gene expression, gene detection, Caenorhabditis elegans, support vector machines, decision trees

Joint work of: Li, Yang; Breitling, Rainer

New algorithms for high-resolution metabolomics - A case study on trypanosome parasites

Rainer Breitling (University of Groningen, NL)

New mass spectrometry technologies, which provide unprecedented mass accuracy and resolution, promise to generate exciting new datasets in metabolomics. We propose new computational methods that enable us to infer metabolic networks from the comprehensive high-accuracy measurements of small molecules in biological samples. We illustrate the potential of these concepts in a case study on the sleeping sickness parasite, Trypanosoma brucei, a small parasite with a well-delimited metabolism. New software tools implementing these approaches are currently under development and will bring the power of high-resolution metabolomics to the wet lab biologist.

Keywords: Metabolomics, Fourier-Transform Ion Cyclotron Resonance Mass Spectrometry, Orbitrap Mass Spectrometry, ab initio metabolic network reconstruction

Joint work of: Breitling, Rainer; Barrett, Michael

"Learning and Classifying Anisotropic 3D morphologies and Structures in Biology and Medicine"

Hans Burkhardt (Universität Freiburg, D)

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 based on monomial and relational kernel functions.

Examples and applications will be given from the following projects:

1. Self-Learning Segmentation and Classification of Tissue Cell-Nuclei in 3D Volumetric Data using Voxel-Wise Gray Scale Invariants
2. Development of a fast Search Engine for Protein Fold Databases based on Invariant 3D Features
3. Automatic Classification of 3D Chromosome Territories of human lymphocyte Nuclei from confocally scanned two or three color FISH data.
4. Automatic Classification of Airborne Pollen-Grains recorded with a Confocal Laser Scanning Microscope.
5. Automatic Segmentation of Dendritic Spines in 3D LSM Data

fMRI hemodynamic response function from multigrid Bayesian analysis

Nestor Caticha (University of Sao Paolo, BR)

We present a non parametric Bayesian multi scale method to characterize the Hemodynamic Response HR as function of time. This is done by extending and adapting the Multigrid Priors (MGP) method proposed in (S.D.R. Amaral, S.R. Rabbani, N. Caticha, Multigrid prior for a Bayesian approach to fMRI, NeuroImage 23 (2004) 654-662; N. Caticha, S.D.R. Amaral, S.R. Rabbani, Multigrid Priors for fMRI time series analysis, AIP Conf. Proc. 735 (2004) 27-34). We choose an initial HR model and apply the MGP method to assign a posterior probability of activity for every pixel. This can be used to construct the map of activity. But it can also be used to construct the posterior averaged time series activity for different regions. This permits defining a new model which is only data dependent. Now in turn it can be used as the model behind a new application of the MGP method to obtain another posterior probability of activity. The method converges in just a few iterations and is quite independent of the original HR model, as long as it contains some information of the activity/rest state of the patient.

We apply this method of HR inference both to simulated and real data of blocks and event-related experiments. Receiver operating characteristic (ROC) curves are used to measure the number of errors with respect to a few hyperparameters. We also study the deterioration of the results for real data, under information loss. This is done by decreasing the signal to noise ratio and also by decreasing the number of images available for analysis and compare the robustness to other methods.

Keywords: FMRI, Bayesian data analysis, multiscale, nonparametric

Joint work of: Caticha, Nestor; Amaral, Selene ; Rabbani, Said

See also: NeuroImage 23 (2004) 654, Amaral et al Neuroimage online 19 dec 2006

Reinforcement Learning for Manifold Identification

Colin Fyfe (University of Paisley, GB)

We use several forms of reinforcement learning to adapt parameters of a nonlinear mapping between the points which lie in structured positions in a latent space to their projections in data space. The result is the identification of non-linear manifolds.

Keywords: Reinforcement learning, manifold identification

Median and Relational Neural Gas

Barbara Hammer (TU Clausthal, D)

Neural Gas and SOM provide efficient topographic map formation, but in the original version only for euclidean data. We present extensions to relational (dissimilarity) data via batch optimization schemes: median optimization which restricts ptotoypes to tha data locations and relational neural gas which allows continuous updates based on the relational dual.

Keywords: Neural gas, dissimilarity data

Joint work of: Hammer, Barbara; Hasenfuss, Alexander

Relational Clustering

Barbara Hammer (TU Clausthal, D)

We introduce relational variants of neural gas, a very efficient and powerful neural clustering algorithm. It is assumed that a similarity or dissimilarity matrix is given which stems from Euclidean distance or dot product, respectively, however, the underlying embedding of points is unknown. In this case, one can equivalently formulate batch optimization in terms of the given similarities or dissimilarities, thus providing a way to transfer batch optimization to relational data. Interestingly, convergence is guaranteed even for general symmetric and nonsingular metrics.

Keywords: Neural gas, dissimilarity data

Joint work of: Hammer, Barbara; Hasenfuss, Alexander

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1118

Discriminative and associative clustering, and data fusion

Samuel Kaski (Helsinki Univ. of Technology, FIN)

I will review our work on combining data sets of co-occurring samples to combat noise and irrelevant variation in the data, and introduce new methods. The learning metrics principle shows how to construct metrics where distances are relevant to auxiliary data, such as class distributions. Discriminative clustering applies the principle to clustering. In effect, the clustering becomes supervised by the classes. Associative clustering is symmetric; two data sets are clustered such that each supervises each other, in total maximizing mutual information. This can be regarded as data fusion which focuses on shared properties in data sets. Finally, I will introduce Bayesian gererative methods for such data fusion, including a generalization to canonical correlation analysis and a clustering algorithm.

Keywords: Associative clustering, discriminative clustering, data fusion

Full Paper:
http://www.cis.hut.fi/projects/mi

Segmentation of dynamic positron emission tomography images using cluster analysis: caveats and useful data preprocessing steps to obtain an accurate tumor delineation

John A. Lee (University of Louvain, B)

Segmentation of dynamic positron emission tomography images using cluster analysis: caveats and useful data preprocessing steps to obtain an accurate tumor delineation

Modern radiotherapy treatment machines are able to deliver dose distributions with high gradients. This allows the physicians to maximize the dose received by the tumor while sparing the surrounding organs at risk. In order to achieve this goal, they need an accurate delineation of the tumor. While they usually delineate manually the target on anatomic images (e.g. computed tomography), this is more difficult for functional imaging modalities like positron emission tomography (PET) because:

1. The spatial resolution of PET is (images are blurred).
2. PET images can be dynamic, i.e. are vector-valued in order to record the uptake of the injected radioactive tracer.

In the case of dynamic images, the time activity curves (TACs) can be clustered in order to obtain a segmentation of the images. This relies on the assumption that the different tissues (tumor, surrounding inflammation, muscle, bones, etc.) have a different behavior in time. Because of the low resolution, however, there can be a large overlap between the underlying clusters. In addition, as the neighborhood relationships between pixels are not taken into account, we can end up with useless results (e.g. concentric clusters near the boundary of the tumor, which is not a satisfactory result). Fortunately, the use of a simple model of PET imaging that accounts for the low resolution and specific noise allows us to derive appropriate image processing tools. With the processed images, the cluster analysis is made easier and the overlap between the different tissue classes is reduced.

Validation with surgical specimen with known tumor volume leads to better results (in terms of volume and mismatch) than those of classical segmentation methods used for PET images.

Joint work of: Lee, John A.; Geets, Xavier; Grégoire, Vincent

Learning the Dependency Structure of Heterogeneous Datasets Using Mixtures of Graphical Gaussian Models

Johannes Mohr (Bernstein Center for Comp. Neuroscience - Berlin, D)

Learning the structure of a Gaussian Graphical Model (GGM) from a given dataset allows to study the (in)dependencies among a set of variables. In biomedical applications, however, often the dataset can be assumed to be quite heterogeneous, being comprised of several groups of cases with different dependency structure. I will show how mixtures of GGMs can be used to model that kind of data, and propose an EM algorithm for learning such models.

This talk is about work still in progress, so only some preliminary but promising results will be shown.

Acknowledgements: This work was funded by the Bernstein Center for Computational Neuroscience Berlin (BMBF grant 01GQ0411).

Dissimilarity Clustering

Fabrice Rossi (INRIA Rocquencourt, F)

This talk presents some improvements to the Median Self Organizing Map (SOM) and a short tutorial on non prototype based clustering for dissimilarity data.

The Median SOM is a simple extension of the Batch SOM for dissimilarity. It gives interesting results but suffers from a important algorithmic cost: for N input data and M neurons, the cost of one iteration if $O(N^2M)$.This talk presents an algorithmic modification that lead to a cost of $O(N^2 + NM^2)$, as well as branch and bound tricks, that reduce the running time from hours to minutes for medium size data (e.g., $N = 3000$ and $M = 300$).

Another limitation of the Median is related to the fact that prototypes are chosen among the data, leading to poor quantization of the input space and to prototype collisions. The talk presents a simple branch and bound way to find an approximation of the optimal prototype lists when collisions are forbidden.

Finally, the talk introduces briefly a clustering method proposed by Buhmann and Hofmann, that minimizes dissimilarity between all members of a cluster, leading to robustness against non metricity in the dissimilarity. A mean field annealing approach can be used to optimized the obtained cost function.

Keywords: Dissimilarity, Clustering, Self Organizing Map, Simulated Annealing, Pairwise data

Inference by message passing in dense and composite systems

David Saad (Aston University - Birmingham, GB)

Probabilistic graphical models provide a powerful framework for modelling statistical dependencies between variables, mainly in systems that can be mapped onto sparse graphs. They play an essential role in providing principled probabilistic inference in a broad range of applications from medical expert systems, to telecommunication.

These methods, that have largely been developed independently in the computer science and information theory literature, also have deep roots in advanced mean field methods of statistical physics.

Message passing techniques are perceived as impractical for densely connected systems due to the computational effort involved and the existence of loops, but can be used in this context by introducing a set of average messages sampled from a Gaussian distribution, whose parameters are updated iteratively. However, this approach fails when the solution space becomes fragmented, for instance, when there is a mismatch between the assumed and true prior information.

We extended this approach to tackle inference problems where no reliable prior information is available, conceptually in a similar way to the extension of belief propagation to survey propagation in the case of sparse graphs, by replicating the system variables and calculating pseudo-posterior estimates based on averages over the replicated systems. This is carried out by considering an infinite number of replicated systems and employing methods of statistical physics. The method has been applied to CDMA signal detection and learning in Ising linear perceptron showing optimal performance for large systems.

The new approach also facilitates the use of message passing for inference in composite systems that comprise different levels of connectivity and interaction strengths. We have successfully applied the approach to toy composite models.

Finally, we will review the application of this inference method to other problems in communication and possible extensions.

Keywords: Inference, message passing

Joint work of: David Saad, Juan P Neirotti and Etienne Mallard

See also: J.P. Neirotti and D. Saad, Europhys. Lett. Vol. 71 866 (2005); J.P. Neirotti and D. Saad, Physica A Vol. 365 203 (2006)

Estimation of boar sperm quality using intracellular distributions and contour-based techniques

Lidia Sanchez-Gonzalez (Univ. of León, E)

We propose several methods in order to assess automatically the viability and the fertility of boar sperm head images.

In the first case, there are two classes: alive and dead sperm cells. We compute a 2D grey level function that represents the intracellular density distribution of such cells and we consider such function to model alive cell images. We obtain a decision criterion to classify a test set formed by different alive and dead cell images.

In the second case, there are also two classes: acrosome-intact and acrosome-damaged spermatozoa. We obtain the boundaries of the sperm head images and compute a set of different contour descriptors in order to discriminate between the two existing classes.

Advances in pre-processing and model generation for mass spectrometric data analysis

Frank Michael Schleif (Universität Leipzig, D)

The analysis of complex signals as obtained by mass spectrometric measurements is complicated and needs an appropriate representation of the data. Thereby the kind of feature extraction as well as the used similarity measure are of particular importance.

Focusing on biomarker analysis and taking the functional nature of the data into account this task is even more complicated. Different approaches are shown and applied in an analysis of data taken from clinical proteom studies.

Keywords: Functional data, signal processing, similarity measure, metric adaptation, biomarker, mass spectrometry, proteomics

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1132

Relevance Matrices in LVQ

Petra Schneider (University of Groningen, NL)

LVQ-networks belong to the class of distance-based classifiers. The underlying distance measure is of special importance for their performance, because it defines how the data items are compared and how they are grouped in clusters.

Relevance Learning techniques try to adapt the distance measure to the specific data used for training. I will present a new adaptive distance measure in Learning Vector Quantization which is an extension of previously proposed Relevance Learning schemes. In comparison to the already existing techniques for Relevance Learning, this distance measure is more powerful to represent the internal structure of the data appropriately.

Two applications will be used to demonstrate the behavior of the new algorithm (artificial and real life).

Keywords: Learning Vector Quantization, Relevance Learning, adaptive distance measure

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1133

Correlation-based Data Representation

Marc, Strickert (IPK Gatersleben)

The Dagstuhl Seminar 'Similarity-based Clustering and its Application to Medicine and Biology' (07131) held in March 25–30, 2007, provided an excellent atmosphere for in-depth discussions about the research frontier of computational methods for relevant applications of biomedical clustering and beyond. We address some highlighted issues about correlation-based data analysis in this seminar postribution. First, some prominent correlation measures are briefly revisited. Then, a focus is put on Pearson correlation, because of its widespread use in biomedical sciences and because of its analytic accessibility. A connection to Euclidean distance of z-score transformed data outlined. Cost function optimization of correlation-based data representation is discussed for which, finally, applications to visualization and clustering of gene expression data are given.

Keywords: Correlation, data representation, gradient-based optimization, clustering, neural gas

Joint work of: Strickert, Marc; Seiffert, Udo

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1134

Linear patient model combining for the detection of Interstitial Lung Disease

David M. J. Tax (TU Delft, NL)

In some classification problems, like the detection of illnesses in patients, classes are very unbalanced and the misclassification costs for different classes vary significantly. Then it is better not to minimize the classification error, but to optimize the ordering of the data, or to optimize the Area under the ROC curve (AUC). In this talk I propose to optimize a linear combination of features (or base model outputs) by optimizing AUC. The advantages are that a relatively small training set is required for the optimization and that the training set can have a large class imbalance. Furthermore, the classifier does not make distributional assumptions, making it very suitable to combine the outputs of base classifiers. In the application of the detection of interstitial lung diseases it is shown to be very advantageous and to outperform standard classification rules.

Keywords: Pattern recognition, ROC curve, area under the ROC curve

On Topographic Maps/Clustering of Structured Data

Peter Tino (University of Birmingham, GB)

I will talk about general principles of extending topographic maps of vectorial data to more complex structured data types. I will then concentrate on generative probabilistic modeling.

The model is basically a constrained mixture of appropriate noise models. The approach will be illustrated on chorals by J.S. Bach and fluxes from eclipsing binary stars.

Learning Vector Quantization: generalization ability and dynamics of competing prototypes

Aree Witoelar (University of Groningen, NL)

Learning Vector Quantization (LVQ) are popular multi-class classification algorithms. Prototypes in an LVQ system represent the typical features of classes in the data. Frequently multiple prototypes are employed for a class to improve the representation of variations within the class and the generalization ability. In this paper, we investigate the dynamics of LVQ in an exact mathematical way, aiming at understanding the influence of the number of prototypes and their assignment to classes. The theory of on-line learning allows a mathematical description of the learning dynamics in model situations. We demonstrate using a system of three prototypes the different behaviors of LVQ systems of multiple prototype and single prototype class representation.

Keywords: Online learning, learning vector quantization

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2007/1131

Full Paper:
http://jmlr.csail.mit.edu/papers/v8/biehl07a.html

See also: Journal of Machine Learning Research (8): 323-360 (2007)

1. A Tutorial on Spectral Clustering; 2. Stability and Resampling Methods for Clustering

Ulrike von Luxburg (MPI für biologische Kybernetik - Tübingen, D)

This tutorial derives spectral clustering by several different approaches, compares different spectral clustering algorithms, and explains their relations them to other clustering algorithms.

Depending on the preference of the audience, I will focus on algorithmic, theoretic, or implementation details.

Stability and Resampling Methods for Clustering

Resampling methods are a very popular tool for tuning parameters of clustering algorithms, in particular for choosing the number k of clusters to construct. However, it seems that at least for large sample size the method does not really do what people expect. I will try to discuss this discrepancy. Rather than presenting final results, his is about work in progress.