Report from Dagstuhl Seminar 11341

Learning in the context of very high dimensional data

Edited by

Michael Biehl[1], Barbara Hammer[2], Erzsébet Merényi[3], Alessandro Sperduti[4], and Thomas Villmann[5]

1 **University of Groningen, NL,** m.biehl@rug.nl
2 **Universität Bielefeld, DE,** bhammer@techfak.uni-bielefeld.de
3 **Rice University, US,** erzsebet@rice.edu
4 **University of Padova, IT,** sperduti@math.unipd.it
5 **Hochschule Mittweida, DE,** thomas.villmann@hs-mittweida.de

> *With real world data it all stands and falls*
> *if knowledge and insight are truly your goals.*
> *But if you like greatly*
> *delicious sweet pastry*
> *you might just as well resort to Swiss Rolls.*
>
> Michael Biehl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 11341 “Learning in the context of very high dimensional data”. The aim of the seminar was to bring together researchers who develop, investigate, or apply machine learning methods for very high dimensional data to advance this important field of research. The focus was be on broadly applicable methods and processing pipelines, which offer efficient solutions for high-dimensional data analysis appropriate for a wide range of application scenarios.

Seminar 22.–26. August, 2011 – www.dagstuhl.de/11341
1998 ACM Subject Classification I.2.6 [Artificial Intelligence] Learning
Keywords and phrases Curse of dimensionality, Dimensionality reduction, Regularization Deep learning, Visualization
Digital Object Identifier 10.4230/DagRep.1.1.67

Learning in the context of very high dimensional data, *Dagstuhl Reports*, Vol. 1, Issue 1, pp. 67–95
Editors: M. Biehl, B. Hammer, E. Merényi, A. Sperduti, and T. Villmann
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Michael Biehl
Barbara Hammer
Erzsébet Merényi
Alessandro Sperduti
Thomas Villmann

Goals of the seminar

Rapidly increasing sensor technology, greatly enhanced storage capabilities, and dedicated data formats have lead to a dramatic growth of the size of electronic data available today and, even more so, its dimensionality. Examples include diverse formats such as spectral data, micro- and macroarrays, biotechnological sequence data, or high resolution digital images. Due to its dimensionality and complexity, these data sets can hardly be addressed by classical statistical methods; nor do standard presentation and visualization tools allow an adequate direct inspection by humans. Thus, the need for efficient and reliable automatic processing and analysis tools for very high dimensional data arises in different areas such as bioinformatics, medicine, multi-band image analysis, robotics, astrophysics, geophysics, etc.

The aim of the seminar was to bring together researchers who develop, investigate, or apply machine learning methods for very high dimensional data to advance this important field of research. The focus was put on broadly applicable methods and processing pipelines which offer efficient solutions for high-dimensional data analysis appropriate for a wide range of application scenarios.

Questions tackled in the seminar included the following areas:

1. **Sparse representation and regularization:**
   a. Which general principles (such as information theory, preservation of inherent data structures) offer suitable frameworks in which to achieve a compact representation of high dimensional data? Is it possible to turn these general principles into efficient algorithmic form as mathematical regularization conditions?
   b. Which models are suitable to represent high dimensional data in a dense form (such as prototype based methods, functional data representation, dedicated algebraic structures, decomposition methods)? What are their adaptive parameters and how can they be adapted?
   c. How can the number of free model parameters be restricted by regularization such that the available data provides a sufficient statistics for the resulting model? Is it possible to derive explicit mathematical bounds on the generalization ability?
2. **Dedicated metrics and kernels for high-dimensional data:**
   a. How can the inherent non-Euclidean structure be inferred from the data in the presence of high dimensionality? Which aspects of particular relevance for the application should be emphasized by the corresponding similarity structure and how can this information be estimated with robust statistical tools?
   b. How can this information be embedded into metrics or kernels? Do there exist particularly suited approaches for high dimensional data of specific form such as kernels which make use of sparsity or functional dependencies of the data? How can this be realized algorithmically in an efficient way regarding the high dimensionality?

   c. Is it possible to partially automate the detection of a suitable similarity structure for the analysis tool and accompany this with guarantees such as consistency, or bounds on the generalization ability in the context of high-dimensionality?

3. **Efficient realizations:**
   a. How can robust learning algorithms be designed for the model parameters, ones that can deal with noise and uncertain data, missing values, etc., particularly pronounced in high dimensional data? Are these techniques insensitive with respect to the choice of the metaparameters (such as learning rate, degree of regularization), such that generic methods suitable for non-experts in the field result?
   b. How can the methods be realized efficiently in the context of very high dimensionality? Methods which are linear in the number of dimensions are probably already too slow in this context.
   c. Can the learning algorithms be realized in such a way that adaptation to new data and life-long learning become possible? What are characteristic time scales required for learning certain parameters in dependence of the data dimensionality?
4. **Evaluation of methods:**
   a. What are inherent evaluation criteria for the reliability of the models, also when the number of data points is small compared to the dimensionality? What are stable conformal predictors for model adequacy and accuracy?
   b. How can the results be presented to experts such that humans can judge the reliability and quality of the model? How can, in turn, user feedback be integrated into the models?
   c. Can simplifying models of learning scenarios give insight into the performance of practical algorithms? Which information visualization methods are suitable in this context?

Structure

39 experts from 12 different countries joined the seminar, including a good mixture of established scientists and promising young researchers working in the field. Thereby, the special interests of the researchers ranged from dedicated algorithmic design connected to diverse areas such as dimensionality reduction, data visualization, metric learning, functional data analysis, to various application scenarios including diverse areas such as the biomedical domain, hyperspectral image analysis, and natural language processing.

This setup allowed us to discuss salient issues in a way that integrated perspectives from several points of views and scientific approaches, thereby providing valuable new insights and research contacts for the participants. Correspondingly, a wide range of topics was covered during discussions and presentations in the seminar.

During the week, 29 talks were presented which addressed different aspects of how to deal with high dimensional data and which can be grouped according to the following topics:

- Dimensionality reduction techniques and evaluation measures
- Biomedical applications
- Distances, metric learning, and non-standard data
- Functional data processing
- Probabilistic models for dimensionality reduction
- Feature selection and sparse representation of data

The talks were divided into a variety of tutorials which gave introductory overviews and opinions on important research directions and shorter talks which focused on specific recent (partially yet unpublished) scientific developments.The talks were supplemented by vivid discussions based on the presented topics as well as the traditional social event on Wednesday afternoon in the form of a visit to the beautiful town of Trier.

Results

A variety of open problems and challenges came up during the week. The following topics were identified as central issues in the context of the seminar:

- **Desired properties of dimensionality reduction techniques, possibilities of their formal evaluation:** The topic of dimensionality reduction and data visualization has been addressed in several presentations including several tutorial talks.
  It became apparent that the topic is currently a very rapidly emerging field in machine learning, with a manifold of advanced algorithms being published in the recent literature. Key issues, however, remain a challenge: to use advanced methods in applications, there is the need for widely parameterless techniques, clear interpretability of the results, and comfortable usage e.g. regarding processing speed or uniqueness and robustness of the results.
  For these reasons, advanced methods are often not used in practice.
  It has been discussed that the desired properties and results of dimensionality reduction techniques depend on the given task at hand and cannot universally be formalized. Nevertheless, there is a need for formal evaluation methods of dimensionality reduction to compare techniques, and to guide parameter choice and optimization.
  Promising general evaluation schemes have been proposed in recent years as presented in the seminar, but an extensive evaluation of their suitability is so far lacking.
- **Good scientific benchmarks and evaluation criteria:** In this context, is has been raised that good, accessible benchmarks are rare. Albeit high dimensional electronic data are ubiquitous, these data often require complex preprocessing, they do not allow evaluation of formal methods due to the lack of objective evaluation information, or they are even sometimes subject to restrictions. For that reason, real life data are partially not accessible, and there is the risk that methods are over-adapted according to the available benchmarks which do not necessarily mirror the demands in practical applications. During the seminar, however, it has been raised that quite a few benchmarks have become available in the context of contest data.
- **Where to use complex models as compared to well-established linear techniques:** During the seminar, it became apparent that there is a gap in between advanced techniques proposed in the context of machine learning for high dimensional data and methods which are actually used in application domains such as biomedical data analysis. Often, in practical applications simple linear techniques seem sufficient to reliably detect important and interpretable information in high dimensional data collections.
  A variety of reasons has been discussed in this context: in particular in bioinformatics, the technology to gather data and large data collections are often comparably novel such that information still lies ‘at the surface’ of the data.
  For very high dimensional data, linear techniques sometimes seem the only methods for which sufficient reliability can be guaranteed, more complex nonlinear methods likely focusing on noise in the data due to the lack of appropriate regularizers which suite

the given setting. In this context, it has been raised that the embedding of data into high-dimensions where linear methods often suffice constitutes one of the most prominent approaches to actually solve standard nonlinear problems, popular examples in this context being the support vector machine, the extreme learning machine, or reservoir computing.

Further, linear techniques seem to focus on 'universally important' issues which are relevant independent of the context due to universal statistical properties. For more advanced techniques, domain knowledge is required to set up the models or to interpret the results in a reliable way. This argument has been substantiated in the seminar by several presentations. For example, knowledge about biological networks and metabolic pathways can be integrated into biological data processing and it can greatly enhance the performance.

- **How to deal with complex structures:** It has been raised that an intelligent preprocessing of the data is often more important than the choice of the model. Alternatively, data can be tackled with appropriate problem adapted metrics, followed by rather simple machine learning techniques. In the seminar, quite a number of complex data formats have been presented in applications, such as e.g. data with inherent functional form connected to spectrometry data. Besides the necessity to come up with suitable metrics, this also leads to very interesting theoretical problems. For example, it can be proved that it is not possible to learn in the space of functions at all, unless very strong requirements are fulfilled.

  Nevertheless, impressive applications have been achieved in this context. Thus, it seems worthwhile to investigate which constraints are fulfilled in practical applications such that learnability is guaranteed. On the other side, a variety of powerful metric adaptation schemes exist ranging from fast and efficient convex techniques to nonlinear cost functions. In most cases, however, a simple Mahalanobis distance is used and, often, only basic machine learnings are integrated such as simple k nearest neighbor.

  Further, a unifying theory and guidelines, which technique is best suited in which scenarios, remain unsolved problems.
- **How to model in the correct way for very high dimensions:** Statistics being the universal background for almost all machine learning techniques, it has been raised that statistical models are often rather uniform in their principled design, not taking advantage of the rather flexible way to model dependencies of data and dimensions. While it is common in the context of simple PCA to swap the role of input dimensions and data points in case of very high dimensions, more complex nonlinear models stick to the classical setting as used in the case of comparably low dimensionality. It could be worthwhile to put different principled modeling paradigms into general patterns which allow to reformulate established techniques such that they become suitable for very high dimensional data.

Altogether, the seminar opened quite a few perspectives pointing into important research directions in the context of very high dimensional data.

2 Table of Contents

Executive Summary

Overview of Talks

Preliminary follow up publications resulting from the seminar

3 Overview of Talks

3.1 Some biology applications for the analytically minded

Gyan Bhanot (Rutgers University – Piscataway, US)

Joint work of Bhanot, Gyan and many others
Main reference Greenbaum B.; Levine AJ., Bhanot G. and Rabadan R., PLoS Pathogens. 2008 Jun 6;
G. Alexe, et al., Can. Res.67, 10669-76, 2007;
Alexe G, et al, 2008, J. Mol Evol, 67 (5), 465-87.

At the seminar, I talked about several successful projects where simple analytical methods and ideas from bioinformatics were able to reveal novel patterns in biological data. Many of these discoveries were validated in wet lab experiments conducted with biologist and medical colleagues and led to novel understanding. I will describe briefly the projects I discussed.

1. The first project was entitled: “Inferring the past” and involved a simple use of PCA and consensus clustering on mitochondrial sequence data to infer the phylogeny of human migration patterns (Alexe G, Vijaya-Satya R, Seiler M, Platt D, Bhanot T, Hui S, Tanaka M, Levine AJ, Bhanot G. PCA and Clustering Reveal Alternate mtDNA Phylogeny of N and M Clades. 2008, J. Mol Evol, 67 (5), 465–487. PMID: 18855041). We found that the tree emerged as a hierarchical pattern of embedded clusters at each level of PCA. Data bootstrapping and consensus clustering was used to compute the robustness of branches and identify the most informative SNPs.
2. The second project was entitled: “PCA, Clustering reveal clinical subtypes in breast cancer”. Here, we used the same methodology to identify clinically relevant classes of breast cancer. The significant discovery here was that HER2+ breast cancers, which are more aggressive, have two subtypes, distinguished by the presence or absence of a lymphocytic infiltration into the tumor which is visible in pathology specimens using a simple staining assay. The clinical utility of this discovery was that the tumors with lymphocytic infiltration were responsive to Herceptin (a drug used in treating HER2+ breast cancers) while those without the infiltrate were less responsive. The paper on this work is :

   G. Alexe, et al. ’High Expression of Lymphocyte Associated Genes in Node Negative HER2+ Breast Cancer correlates with lower Recurrence rates.’ Cancer Research, 67, 10669–10676, 2007.

3 The third project was entitled: “Evolution and Mimicry in Influenza and Other RNA Viruses” which was based on:Greenbaum B, Levine AJ, Bhanot G and Rabadan R, PLoS Pathogens. 2008 Jun 6;4(6):e1000079. In this paper, we used a simple permutation method on CpG in the 3rd and 1st coding positions on the viral genome to identify significant di-nucleotide patterns under selection in Flu. We showed that the 1918 Flu virus (H1N1) has evolved to lower CpG content to make it less visible to the immune system. Our research suggested that specific Toll like receptors must exist which can identify ssRNA in cells. This prediction was subsequently validated by experiments. We were also able to characterize the virulence of emerging strains based on the CpG content and flanking sequence context.

3.2 Admire LVQ: The DREAM6 AML prediction challenge

Michael Biehl (University of Groningen, NL)

Joint work of Biehl, Michael; Bunte, Kerstin; Schneider, Petra

We briefly present and discuss our entry to the DREAM6 AML prediction challenge, 2011.

The construction of feature vectors from the flow cytometry counts based on statistical moments is briefly described. Generalized Matrix Relevance Learning is used to obtain a classification scheme and evaluated with respect to ROC performance. The obtained relevance profile provides further insight into the data and the nature of the problem.

This brief presentation is meant as an addendum to the talk of Marc Strickert, who presents an alternative approach to the problem.

3.3 Adaptive Matrices for Color Texture Classification

Kerstin Bunte (University of Groningen, NL)

Joint work of Bunte, Kerstin; Giotis, Ioannis; Petkov, Nicolai; Biehl, Michael
Main reference Kerstin Bunte, Ioannis Giotis, Nicolai Petkov and Michael Biehl, "Adaptive Matrices for Color Texture Classification, Computer Analysis of Images and Patterns," 14th International Conference, CAIP 2011, Seville, Spain, pp. 489–497.
URL http://dx.doi.org/10.1007/978-3-642-23678-5_58

In this paper we introduce an integrative approach towards color texture classification learned by a supervised framework. Our approach is based on the Generalized Learning Vector Quantization (GLVQ), extended by an adaptive distance measure which is defined in the Fourier domain and 2D Gabor filters. We evaluate the proposed technique on a set of color texture images and compare results with those achieved by simple gray value transformation on the color images with a comparable dissimilarity measure and the same filter bank. The features learned by GLVQ improve classification accuracy and they generalize much better for evaluation data previously unknown to the system [1].

References

1 Kerstin Bunte, Ioannis Giotis, Nicolai Petkov, and Michael Biehl. *Adaptive Matrices for Color Texture Classification*. Computer Analysis of Images and Patterns – 14th International Conference, CAIP 2011, Seville, Spain, August 29–31, 2011, Proceedings, Part II

3.4 Challenges of feature representation in Natural Language Processing tasks

Richard Farkas (Universität Stuttgart, DE)

Natural Language Processing (NLP) tasks are usually traced back to classification or ranking problems. The feature representation of instances (which are words, sentences or documents) usually consists of several millions of features. The features set is heterogeneous, the features can be binary (e.g. occurrence of a particular word) or continuous (e.g. information theoretic measures calculated on unlabeled large document sets); they are usually highly inter-dependent; they are sparse. In spite of these special characteristics, the NLP community employs simple and standard machine learning techniques to handle the feature space. In this talk I focus on the challenges of feature representation in NLP tasks and introduce some promising (task-dependent) approaches to handle the high-dimensionality of these problems.

3.5 Optimization of Parametrized Divergences in Fuzzy c-Means

Tina Geweniger (University of Groningen, NL)

Joint work of Geweniger, Tina; Kästner, Marika; Villmann, Thomas

In many scientific fields like biology, medicine, geology etc. clustering of data plays an important role. Sets of multi-dimensional data samples are grouped to detect or visualize the underlying structure. One family of algorithms are prototype based methods, where each prototype represents one cluster center.

Famous representatives are c-means and neural gas. In my talk I focused on the fuzzy c- means as a variant of the standard c-Means algorithm determining fuzzy memberships to the cluster centers. Usually the Euclidean distance is applied to calculate distances between prototypes and data samples. Yet, if these samples represent functions, i.e. high-dimensional data vectors with spatially correlated components, generalized divergences could be a more appropriate distance measures. Furthermore, incorporation of relevance learning, i.e. weighting of function intervals, leads to improved clustering solutions. We modified the fuzzy c-means such, that the two concepts - divergences and relevance learning - are integrated and presented the theory in combination with some examples. To compare the respective results different cluster evaluation measures for fuzzy clustering were applied and discussed briefly.

3.6 Kernel t-SNE

Andrej Giesbrecht (Universität Bielefeld, DE)

Joint work of Giesbrecht, Andrej; Lueks, Wounter; Hammer, Barbara

The visualization of high-dimensional data is an important challenge in the current data mining research field. A low-dimensional representation of data (e.g. 2D) is an accessible and intuitive way for humans to analyze the information hidden in the complex data. For this reason many algorithms emerged recently to address this problem. However most of these techniques have quadratic or even cubic complexity. One way to overcome this problem is to train the mapping on a manageable subset of the data and then project the remaining data using this mapping. In our contribution we investigate this approach using t-distributed Stochastic Neighbor Embedding [1], which is one of the most well-known methods in the area. We compare the performance of the out-of-sample extension of t-SNE, which is based on the gradient descend, with a simple interpolation technique and SVM regression. All the three methods are using the result of the trained t-SNE mapping. Also we propose a new method called kernel t-SNE, which is based on t-SNE. Instead of computing the low-dimensional representation of the training data, it learns a kernel mapping, which can be directly applied to the new data. The results show that the interpolation technique and kernel t-SNE achieve good performance, comparable to the out-of-sample extension of t-SNE, being faster than the latter.

References

1 L. van der Maaten and G. Hinton. *Visualizing data using t-sne*. Journal of Machine Learning Research, vol. 9, pp. 2579–2605, November 2008.

3.7 Functional Relevance Learning in Generalized Learning Vector Quantization

Marika Kästner (Hochschule Mittweida, DE)

Joint work of Kästner, Marika; Villmann, Thomas
Main reference M. Kästner, T. Villmann, "Functional Relevance Learning in Learning Vector Quantization for Hyperspectral Data," Proceedings of WHISPERS 2011.

Classification accuracy of functional data frequently depends on only a few dimension windows distinguishing different classes. Hence, classifier systems should not only achieve a good performance but also figure out what is essential for this decision. Learning vector quantization is a robust prototypebased classification method which, together with the relevance learning strategy, assesses the relative contribution of spectral bands for efficient classification. Original relevance learning is based on the scaled Euclidean distance, weighting each band independently. This yields a vectorial relevance profile indicating those dimensions which distinguish the classes best. We propose the use of the functional Sobolev distance instead of the Euclidean, together with a relevance function as profile taking into account the functional properties of data. The relevance function is a superposition of a small set of simple basis functions like Gaussians or Lorentzians. In this way the number of parameters to be optimized in relevance learning is drastically decreased such that an inherent stabilization is obtained while the classification accuracy level is retained. We demonstrate the ability of the functional approach for ground cover classification of an AVIRIS hyperspectral data set (Lunar Crater Volcanic Field). In particular it is emphasize model sparsity in terms of structural sparsity and feature selection.

References

1 M. Kästner, B. Hammer & T. Villmann. Generalized Functional Relevance Learning Vector Quantization. In M.Verleysen (Edt.) *Proc. of European Symposium on Artificial Neural Networks (ESANN'2011)*, pages 93.98, Evere, Belgium, 2011.

2 M. Mendenhall and E. Merényi. Relevance-based feature extraction for hyperspectral images *IEEE Transactions on Neural Networks*, 19(4), 658-672, 2008.

3 B. Hammer & T. Villmann. Generalized Relevance Learning Vector Quantization. *Neural Networks*, 15,1059-1068, 2002.

4 A. S. Sato & K. Yamada. Generalized Learning Vector Quantization *Advanced in neural information processing systems*, 7, 423-429,1995.

3.8 Generative modeling of dependencies between high-dimensional data sets

Arto Klami (Aalto University, FI)

Canonical correlation analysis (CCA) finds maximally correlating linear components of two data sets. Formulated as explicit maximization of correlation, the model severely overfits to small sample sizes and high-dimensional data.

Solving the same problem via a Bayesian formulation helps for small sample sizes, but the associated generative model cannot be reliably estimated for high-dimensional data, severely limiting the applicability of Bayesian CCA. In this talk we show how Bayesian CCA can be used also for high-dimensional data by re-formulating the model as a simpler matrix factorization with group-wise sparsity structure. We present an efficient variational Bayesian algorithm for inference, provide a new kind of multi-set CCA extension, and demonstrate the model in analysis of, e.g., high-dimensional brain imaging data.

3.9 Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants

John A. Lee (University of Louvain, BE)

Joint work of Lee, John A.; Verleysen, Michel;

Dimensionality reduction aims at representing high-dimensional data in low-dimensional spaces, mainly for visualization and exploratory purposes. As an alternative to projections on linear subspaces, nonlinear dimensionality reduction, also known as manifold learning, can provide data representations that preserve structural properties such as pairwise distances or local neighborhoods. Very recently, similarity preservation emerged as a new paradigm for dimensionality reduction, with methods such as stochastic neighbor embedding and its variants. Experimentally, these methods significantly outperform the more classical methods based on distance or transformed distance preservation.

This talk explains both theoretically and experimentally the reasons for these performances. In particular, it details (i) why the phenonomenon of distance concentration is an impediment towards efficient dimensionality reduction and (ii) how SNE and its variants circumvent this difficulty by using similarities that are invariant to shifts with respect to squared distances. The talk also proposes a generalized definition of shift-invariant similarities that extend the applicability of stochastic neighbour embedding to noisy data.

3.10 Scale-independent quality criteria for dimensionality reduction

John A. Lee (University of Louvain, BE)

Joint work of Lee, John A.; Verleysen, Michel;

Dimensionality reduction aims at representing high-dimensional data in low-dimensional spaces, in order to facilitate their visual interpretation. Many techniques exist, ranging from simple linear projections to more complex nonlinear transformations. The large variety of methods emphasizes the need of quality criteria that allow for fair comparisons between them. This talk extends previous work about rank-based quality criteria and proposes to circumvent their scale dependency. Most dimensionality reduction techniques indeed rely on a scale parameter that distinguishes between local and global data properties. Such a scale dependency can be similarly found in usual quality criteria: they assess the embedding quality on a certain scale. Experiments with various dimensionality reduction techniques eventually show the strengths and weaknesses of the proposed scale-independent criteria.

3.11 Multiple Instance Learning

Marco Loog (TU Delft, NL)

Joint work of Loog, Marco; Tax, David M.J.; Sørensen, Lauge; Cheplygina, Veronika; Lee, Wan-Jui; Duin, Robert P.W.;

We present multiple instance learning, or rather multiset classification, as a technique that can play a key role in modeling high dimensional, complicated classification tasks. Where in the standard classification every object is described by a single feature vector, in the multiset classification setting, every object is represented by a collection, a multiset, of feature vectors.

The number of feature vectors may differ from object to object. Example problems where a single feature vector typically does not suffice, but certain object parts can be robustly characterized by feature vectors, are web page classification, document labeling, movie rating, music classification, and gesture recognition. It is classically applied to the problem of molecule classification, but nowadays many of its applications can be found within the fields of computer vision, medical image analysis, and computer-aided diagnosis.

As an illustration, we sketch a complete, yet basic, medical image classification pipeline. Going through various dimensionality reduction steps, our original four million dimensional problem is reduced to a multiset classification problem, which dimensionality is in the order of tens. A condensed overview of different approaches to multiset classification is provided. We advocate the use of fusion-based and dissimilarity-based approach to multiset classification, both of which rely on standard classification methods, avoiding special purpose multiple instance learning routines.

References

1 T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

2 M. Loog and B. van Ginneken. Static posterior probability fusion for signal detection: applications in the detection of interstitial diseases in chest radiographs. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 1, pages 644–647, 2004.

3 O. Maron and A.L. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the 15th International Conference on Machine Learning*, 1998.

4 S. Ray and M. Craven. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 697–704, 2005.

5 L. Sørensen, M. Loog, D. Tax, W.J. Lee, M. de Bruijne, and R. Duin. Dissimilarity-based multiple instance learning. In *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS, pages 129–138. Springer, 2010.

6 D. Tax and R. Duin. Learning curves for the analysis of multiple instance classifiers. In *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS, pages 724–733. Springer, 2008.

7 D.M.J. Tax, M. Loog, R.P.W. Duin, V. Cheplygina, and W.-J. Lee. Bag dissimilarities for multiple instance learning. In *1st SIMBAD Workshop*, LNCS. Springer, 2011.

3.12 On the Problem of Finding the Least Number of Features by L1-Norm Minimisation

Thomas Martinetz (Universität Lübeck, DE)

Joint work of Martinetz, Thomas; Klement, Sascha
Main reference Sascha Klement, Thomas Martinetz, "A new approach to classification with the least number of features," ICMLA 2010, Washington, D.C, USA, 12–14 December, 2010, IEEE Computer Society, pp. 141–146.

We proposed the so-called Support Feature Machine (SFM) as a novel approach to feature selection for classification. It relies on approximating the zero-norm minimising weight vector of a separating hyperplane by optimising for its one-norm. In contrast to the L1-SVM it uses an additional constraint based on the average of data points.

In experiments on artificial datasets we observe that the SFM is highly superior in returning a lower number of features and a larger percentage of truly relevant features. Here, we derive a necessary condition that the zero-norm and 1-norm solution coincide. Based on this condition the superiority can be made plausible.

3.13 How to Evaluate Dimensionality Reduction? - Improving the Co-ranking Matrix

Bassam Mokbel (Universität Bielefeld, DE)

In order to make very high-dimensional data accessible for visual inspection and exploration, dimensionality reduction tools can embed the data points in a low-dimensional space, and thereby produce a visualization, e.g., in the Euclidean plane. The existing methods for dimensionality reduction all have unique characteristics and favor certain data properties, often these characteristics are controlled via rather unintuitive parameters. However, since the general problem is ill-posed, it is unclear which is the best embedding solution for a given visualization task. Recently, this has inspired the development of quality assessment measures, in order to evaluate visualization results independently from the methods' inherent criteria. Several quality measures can be (re)formulated based on the so-called co-ranking matrix, which subsumes all rank errors, i.e., differences between the ranking of distances from every point to all others, comparing the low-dimensional representation to the original data. Some measures use a parameter K to divide the co-ranking matrix at the K-th row and column into rectangular submatrices, calculating weighted combinations from the sums of each submatrix's elements. The evaluation process typically involves plotting a graph over several (or even all possible) settings of K. Considering simple artificial examples, we argue that this parameter controls two notions at once, that need not necessarily be combined, and that the rectangular shape of submatrices is disadvantageous for an intuitive interpretation of the parameter. We debate that quality measures, as general and flexible evaluation tools, should have parameters with a direct and intuitive interpretation as to which specific error types are tolerated or penalized for a particular visualization task. Therefore, we propose to replace the parameter K with two distinct parameters to control these notions separately, and introduce a differently shaped weighting scheme on the co-ranking matrix. The two new parameters can then directly be interpreted as a threshold up to which rank errors are tolerated, and a threshold up to which the rank-distances are significant for the quality evaluation. Moreover, we propose a color representation of local quality to support the evaluation process for a given mapping, where every point is colored according to its local contribution to the overall quality value.

3.14 DLVQ and its application to Crop Surveillance

Ernest Mwebaze (Makerere University – Kampala, SAF)

Joint work of Mwebaze, Ernest; Biehl, Michael; Quinn, John A.

DLVQ is a variant of LVQ that uses divergences in the distance measure. This is applicable for non-negative usually normalized data for example histograms and spectra data. We use DLVQ with increasingly complicated 'distance' formulations specified by a combination of partial distances related to more than one type of data. We apply this to Crop Surveillance by representing cassava plant leaf images as histograms that are combined in a compound distance and trained using DLVQ. We implement this classifier on $100 Android mobile phones for automated visual based classification. We also propose dimensionality reduction using causal analysis and present some on-going problems and datasets.

3.15 Is the k-NN classifier in high dimensions affected by the curse of dimensionality?

Vladimir Pestov (University of Ottawa, CA)

Main reference Vladimir Pestov, "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?" arXiv:1110.4347v1 [stat.ML]
URL http://arxiv.org/abs/1110.4347

There is an increasing body of evidence suggesting that exact nearest neighbour search in high-dimensional spaces is affected by the curse of dimensionality at a fundamental level. Does it necessarily mean that the same is true for k nearest neighbours based learning algorithms such as the k-NN classifier? We analyse this question at a number of levels and show that the answer is different at every layer that we peel. As our first main result, we show the consistency of a k approximate nearest neighbour classifier. However, the performance of the classifier in very high dimensions is provably unstable.

As our second major result, we point out that the existing model for statistical learning is oblivious of dimension of the domain and so every learning problem admits a universally consistent reduction to the one-dimensional case.

3.16 Bongard problems: learning in unlimited feature space

John Quinn (Makerere University – Kampala, SAF)

In machine learning, some or all aspects of the type of model and features used for any particular problem are usually selected through human judgment. It is interesting to consider the issues that would need to be solved in order to make entirely automated learning possible. In this talk I discuss Bongard problems, the essence of which is to find classification rules in a setting where the representation cannot be fixed a priori. The need to search for the right representation makes some of the issues in fully automated learning explicit.

3.17 Functional data analysis and learnability in arbitrary spaces

Fabrice Rossi (TELECOM-Paristech – Paris, FR)

I will summarize in this talk recent results about learning in infinite dimensional spaces and more generally in arbitrary spaces. The main motivation for studying this theoretical problem is given by functional data analysis: in this context, each observation point is a function and the goal is to learn e.g., to classify those functions. In practice, one might one for instance to detect some particular compound based on a near infrared spectrum of samples under study.

In practice, functional data are obtained as high dimensional vectors through a natural sampling strategy. For instance a function f is given by $(f(t_1), ..., f(t_d))$. This raises two questions: 1) can we learn to classify exact functions (that is assuming that each function f is completely known) based on a finite learning set? 2) can we learn to classify exact functions based on a finite learning set of sampled functions?

I will first recall that many machine learning algorithms do not need strong assumptions on the input space, at least to be "conceptually" implemented. For instance the K nearest neighbors (KNN) algorithm uses only a dissimilarity on the data and can therefore be applied to any metric space. Linear models can defined simply in Hilbert spaces and then extended to nonlinear mapping using the classical multi-layer perceptron trick (see [1]).

Then I will introduce impossibility examples from [2] and [3]. Those papers show that KNN is not consistent in arbitrary metric spaces and therefore that it cannot be used to learn in this context. [4] shows that the classical kernel estimator is also non consistent in arbitrary metric spaces.

I will give conditions on the space that bring consistency back to KNN. Firstly it is well known since Cover and Hart [5] that the metric space has to be separable (http://en.wikipedia.org/wiki/Separable_space). This is a rather mild condition for functional spaces, but some important spaces such as the one of functions with bounded variations are not separable. Secondly, we need a complex condition that involves both the distribution of the data points and regression function, the so called Besicovitch condition, detailed in [2]. While this condition is automatically true in R^d, it is not in infinite dimensional spaces. Strong assumptions on the regression function (for instance continuity) and/or on the distribution of the data points (for instance a Gaussian distribution with eigenvalues that decrease exponentially quickly) are needed to ensure the consistency of the KNN rule in separable metric spaces of infinite dimension, i.e., in functional spaces considered in functional data analysis.

I will conclude the talk by briefly explaining how one can obtain consistency for functional data analysis with sampled functions. The main idea is to consider regular functions, for instance a Sobolev space such as H^2 (http://en.wikipedia.org/wiki/Sobolev_space). Then natural hypotheses can be used to obtain consistency, for instance a uniform bound on the second derivatives of the data points. Details can be found in [6]. Another solution consists in using projection of the data points on the first functions of a Hilbert basis, as described in [7] and [8].

References

1 Fabrice Rossi and Brieuc Conan-Guez *Functional Multi-Layer Perceptron: a Nonlinear Tool for Functional Data Analysis.* Neural Networks, volume 18, number 1, pages 45-60. January 2005. http://fr.arxiv.org/abs/0709.3642v1

2 F. Cérou and A. Guyader *Nearest Neighbor Classification in Infinite Dimension.* Neural Networks, volume 18, number 1, pages 45-60. January 2005. http://fr.arxiv.org/abs/0709.3642v1

3 Biau, G., Cérou, F. and Guyader, A. *Rates of convergence of the functional k-nearest neighbor estimate.*IEEE Transactions on Information Theory, Vol. 56, pp. 2034–2040. 2010. http://www.lsta.upmc.fr/BIAU/bcg.pdf

4 Abraham, C., Biau, G. and Cadre, B. *On the kernel rule for function classification, Annals of the Institute of Statistical Mathematics.* Vol. 58, pp. 619–633. 2006. http://www.lsta.upmc.fr/BIAU/abc5.p

5 Fabrice Rossi and Nathalie Villa-Vialaneix *Nearest Neighbor Pattern Classification.* T.M. Cover and P.E. Hart. IEEE Transactions on Information Theory, IT-13(1):21–27, January 1967. http://www.stanford.edu/~cover/papers/transIT/0021cove.pdf

6 *Consistency of Functional Learning Methods Based on Derivatives.* Pattern Recognition Letters, volume 32, number 8, pages 1197-1209. June 2011. http://fr.arxiv.org/abs/1105.0204

7 Biau, G., Bunea, F. and Wegkamp, M.H *Functional classification in Hilbert Spaces.* IEEE Transactions on Information Theory, Vol. 51, pp. 2163–2172. 2005. http://www.lsta.upmc.fr/BIAU/bbw.ps

8 Fabrice Rossi and Nathalie Villa-Vialaneix *Support Vector Machine For Functional Data Classification.* Neurocomputing, volume 69, number 7-9, pages 730-742.

3.18 Supervised learning of short and high-dimensional temporal sequences

Frank-Michael Schleif (Universität Bielefeld, DE)

Joint work of Schleif, Frank-Michael; Hammer, Barbara; Gisbrecht, Andrej

Temporal data, with many measurement points and only one or few variables are common in different applications and have been widely studied in the last years. The analysis of short temporal data with many variables is however a quite new field of research and mostly focused on unsupervised approaches like temporal clustering techniques. Here we review first approaches for such data and present a method for the supervised analysis of short and high-dimensional temporal data. The method is evaluated for different artificial data sets.

3.19 Acquisition and processing of high-dimensional data by means of hyperspectral imaging

Udo Seiffert (Fraunhofer IFF Magdeburg, DE)

Hyperspectral imaging is an evolving technology that provides a quantitative assessment of the molecular composition of a wide range of different samples in a non-invasive (optical) manner. The result of the image acquisition process is a stack of images containing the local distribution of reflection spectra of the recorded scenery. Each image pixel becomes a vector along the acquired wavelength range. In contrast to standard colour imaging or multispectral imaging this vector represents a dense and equidistant sampling of a wide wavelength range. This typically leads to a number of characteristic properties of hyperspectral data:

1. High-dimensionality, typically several hundred dimensions;
2. Individual positions of these vectors are rather sampling points of complex patterns than more or less independent features, leading to patterns instead of feature sets;
3. Extensive and proportionate sampling (large number of patterns) due to spatial resolution.

Machine learning offers a powerful framework for pattern recognition and statistical modelling within this context. In order to derive meaningful information from hyperspectral data and comprehensively exploit this imaging technology, novel and adapted data processing approaches and particularly learning paradigms are desired. Hence, hyperspectral image analysis offers both novel perspectives in an increasingly wide range of real-world applications and a challenging playground to develop and test novel methods in machine learning and beyond.

3.20 Functional MRI Analysis

Diego Sona (Fondazione Bruno Kessler – Trento, IT)

Joint work of Sona, Diego; Avesani, Paolo
Main reference Diego Sona, Paolo Avesani, "Feature Rating by Random Subspaces for Functional Brain Mapping," International Conference on Brain Informatics (BI), LNCS Vol. 6334, pp. 112–123, 2010.
URL http://dx.doi.org/10.1007/978-3-642-15314-3_11

Functional magnetic resonance imaging produces datasets presenting high dimensionality in the feature space, and low dimensionality in the sample space.

In the neuroscience community this issue is addressed with a simple and powerful univariate analysis approach, allowing to find areas in the brain maximally activated by specific cognitive or perceptual tasks. We may refer this analysis as "brain mapping". This approach however limits the retrieval of dependencies between the variables. The application of machine learning (ML) may solve this problem, however, two other difficulties arise. On one side ML models usually suffer the curse of dimensionality. On the other side, the experts always require maps indicating the areas of the brain relevant for the investigated cognitive function. In ML terms this corresponds to the set of features allowing a classifier to have good performance. For this reason we need models able to cope with high dimensionality and exhibiting the grouping effect, i.e., similar features must have similar relevance in the trained models.

This would allow to retrieve all relevant features (also the redundant ones) with a posterior analysis of the trained models.

3.21 Correlative matrix mapping connects high-dimensional data sets

Marc Strickert (Universität Siegen, DE)

Many tasks of data analysis concern the linking of high-dimensional data and their labels or, more generally, their multivariate regression targets.

Canonical correlation analysis is a common approach for mapping both vector spaces in a linear fashion into latent spaces where correlation is optimized.

The traditional approach often fails if the number of data points is smaller than the number of data or label dimensions. Furthermore, high-dimensional data, such as spectral measurements, are often redundant and may undergo transformations, while regression targets like metabolite concentrations are often acquired with greater efforts and should be kept constant.

Correlative matrix mapping (CMM) is an alternative formulation based on learning metrics for directed linear transformation from data to targets. Overfitting is reduced by integrating pairwise data relationships into the mapping: the pairwise distances of data transformed by a metric induced by quadratic forms is aimed to be in maximum correlation with the pairwise distances between the regression targets. Second-order learning, l-BFGS, is well-suited to optimize the required mapping parameters. Matrix ranks less than four related to the quadratic form can be used for directly visualizing the transformed discriminative data space. For further reducing overfitting in the CMM model, a very shallow network approach to sub-linear modeling by k-means clustering of the optimized matrix parameters can be considered. CMM can faithfully address supervised visualizations of classification and regression problems, and it can also act for auto-association, that is, as alternative to principal component mappings. Because of the structural simplicity, optimized model parameters can be interpreted as (pairwise) attribute contributions of input data vectors.

Applications to the identification of molecular descriptors and of relevant document terms are provided in the references, and a MATLAB/GNU-Octave implementation is available at https://mloss.org as package CMM.

References

1 Axel J. Soto and Gustavo E. Vazquez and Marc Strickert and Ignacio Ponzoni. *Target-driven subspace mapping methods and their applicability domain estimation.* Molecular Informatics, 2011

2 Strickert, M.; Soto, A. J. & Vazquez, G. E.; Verleysen, M. (Ed.). *Adaptive matrix distances aiming at optimum regression subspace.*
European Symposium on Artificial Neural Networks (ESANN), D-facto Publications, 2010, 93-98

3 Soto, A. J.; Strickert, M.; Vazquez, G. E. & Milios, E. Butz, C. & Lingras, P. (Eds.). *Subspace Mapping of Noisy Text Documents.* Lecture Notes in Artificial Intelligence, Springer-Verlag Berlin Heidelberg, 2011, LNCS 6657, 377-383

3.22 Verification of Cluster Structure: Escalation of Need and Difficulty for Real, High-Dimensional Data, and Recent Developments

Kadim Taşdemir (EC Joint Research Centre – Ispra, IT) and Erzsébet Merényi (Rice University, US)

Joint work of Taşdemir, Kadim; Merényi, Erzsébet
Main reference Taşdemir, K., Merényi, E, "A Validity Index for Prototype-Based Clustering of Data Sets With Complex Cluster Structures," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 41(4), 1039–1053.

The purpose of this work is evaluation of unsupervised clustering results without reference data, i.e., quantification of how well the clusters returned by an algorithm fit the true data partitions. This is a fundamental challenge of clustering because the data structure and the number of clusters are unknown a priori. Cluster validity indices are commonly used tools for relative cluster validation: for ranking the quality of different clustering results. Real-world applications are increasingly dependent on automatic clustering algorithms for finding intricate structure in high-dimensional, large, complicated data spaces, consequently need reliable validation of the discovered structure. Since labeled reference data for problems with many clusters (let alone unexpected clusters) is rarely available, one has to increasingly turn to cluster validity indices. Many existing indices work well for simple data (clusters are well separated or have parametrical shapes or distributions). Most indices, however, do not work well for data sets with complicated cluster structure (variety of clusters of different shapes, sizes, densities, or overlaps), for one or more of the following reasons: The measures of within-cluster scatter and between-clusters separation - which are combined in various ways in the index formulae - are in most cases based on metric distances, thus directly depend on dimensionality.

They involve large numbers of pair wise distances, which results in poor scaling properties. Many use parametric assumptions, or work with extremes, consequently cannot handle irregular cluster structure.

To alleviate some of these problems, we present Conn_Index (Taşdemir & Merényi, 2009, 2011). It works with a connectivity-based similarity measure, CONN, derived from local density distribution and thus possesses considerable immunity to the curse of dimensionality. It is a prototype- based index, i.e., works only for prototype-based clustering (but for any Vector Quantization prototypes). A positive consequence is that it scales well (the number of pair wise distances that need to be computed scales linearly with the number of data points). Since the scatter and separation measures are density-based Conn_Index handles irregular structure quite well. Given that prototype-based clustering has significant performance advantages for huge data sets, development of prototype-based validity indices is strongly motivated. We demonstrate the superior performance of Conn_Index on simple synthetic data, on some of the UCI benchmark machine learning data sets, as well as on real hyperspectral images with complex cluster structure.

3.23 Topographic Mapping and Dimensionality Reduction of Binary Tensor Data of Arbitrary Rank

Peter Tino (University of Birmingham, GB)

Main reference J. Magut, P. Tino, M. Boden, H. Yan. "Multilinear Decomposition and Topographic Mapping of Binary Tensors," in Artificial Neural Networks (ICANN 2010), pp. 317–326, Lecture Notes in Computer Science, Springer-Verlag, LNCS 6352, 2010.
URL http://www.cs.bham.ac.uk/~pxt/TALKS/bin_tensor_talk.pdf

Current data processing tasks often involve manipulation of multi-dimensional objects - tensors. In many real world applications such as gait recognition, document analysis or graph mining (with graphs represented by adjacency tensors), the tensors can be constrained to binary values only. To the best of our knowledge at present there is no principled systematic framework for topographic maps and dimensionality reduction through decomposition of binary tensors. We propose to achieve this through a generalized multi-linear model for binary tensors.

In the model formulation, to account for binary nature of the data, each tensor element is modeled by a Bernoulli noise distribution. To extract the dominant trends in the data, we constrain the natural parameters of the Bernoulli distributions to lie in a sub-space spanned by a reduced set of basis tensors. Bernoulli distribution is a member of exponential family with helpful analytical properties that allow us to derive an iterative scheme for estimation of the basis tensors and other model parameters via maximum likelihood.

We evaluate and compare the proposed technique with existing real-valued tensor decomposition methods in two scenarios: (1) in a series of controlled experiments involving synthetic data; (2) on a real world biological dataset of DNA sub-sequences from different functional regions, with sequences represented by binary tensors.

3.24 Bayesian Models for Variable Selection that Incorporate Biological Information

Marina Vanucci (Rice University - Houston, US)

Main reference Stingo, F.C., Chen Y.A., Tadesse, M.G. and Vannucci, M., “Incorporating Biological Information into Linear Models: A Bayesian Approach to the Selection of Pathways and Genes,” Annals of Applied Statistics, 2011, accepted.
URL http://www.stat.rice.edu/~marina/publications.html

The analysis of the high-dimensional genomic data generated by modern technologies, such as DNA microarrays, poses challenges to standard statistical methods. In this talk I will describe how Bayesian methodologies can be successfully employed in the analysis of such data. I will look at linear models that relate a phenotypic response to gene expression data and employ variable selection methods for the identification of the predictive genes. The vast amount of biological knowledge accumulated over the years has allowed researchers to identify various biochemical interactions and define different families of pathways. I will show how such information can be incorporated into the model for the identification of pathways and pathway elements involved in particular biological processes.

3.25 A brief tutorial on (linear) Distance Metric Learning

Kilian Weinberger (Washington University, US)

One of the fundamental questions of machine learning is how to compare examples. If an algorithm could perfectly determine whether two examples were semantically similar or dissimilar, most subsequent machine learning tasks would become trivial. For example, in classification settings, one would only require one labeled example per class and could then, during test-time, categorize all similar examples with the same class-label. An analogous reduction applies to regression if a continuous estimate of the degree of similarity were available.

It is not surprising that many popular machine learning algorithms, such as Support Vector Machines, Gaussian Processes, kernel regression , k-means or k-nearest neighbors (kNN) fundamentally rely on a representation of the input data for which a reliable, although not perfect, measure of dissimilarity is known. A common choice of dissimilarity measure is an uninformed norm, like the Euclidean distance. Here it is assumed that the features are represented in a Euclidean subspace in which similar inputs are close and dissimilar inputs are far away. Although the Euclidean distance is convenient and intuitive, it ignores the fact that the semantic meaning of "similarity" is inherently task- and data- dependent. Often, domain experts adjust the feature representations by hand - but clearly, this is not a robust approach. It is therefore desirable to learn the metric (or data representation) explicitly for each specific application.

Guided by this motivation, a surge of recent research has focused on learning metrics for the kNN classification (or regression) rule.

Learning a metric instead of (or in addition to) a classifier can have substantially different implications than learning only the classifier directly. For example, in contrast to most classifiers, metrics can generalize to class categories that were unknown during training time. In this tutorial I review a series of recently published algorithm that learn a Mahalanobis metric explicitly from the data. I highlight general trends and outline future directions of metric learning as a research field.

3.26 Relational Extensions of Learning Vector Quantization

Xibin Zhu (Universität Bielefeld, DE)

Joint work of Zhu, Xibin; Schleif, Frank-Michael; Hammer, Barbara

Prototype-based learning algorithms represent given data in a (sparse) way by means of prototypes, they form decisions based on the similarity of data to prototypes, and training is very intuitive based on Hebbian principles. In addition, prototype-base models have excellent generalization ability, and prototypes offer a compact representation of data which can be beneficial for life-long learning.

Unsupervised prototype-based learning such as k-means, topographic mapping, neural gas, or self-organizing map infer prototypes based on input data only.

Supervised techniques incorporate additionally class labels and try to form class boundaries describing priorly known class labels. One of the most popular techniques in this context is learning vector quantization (LVQ), and extensions thereof which are derived from explicit cost function, generalized LVQ (GLVQ), or statistical models, robust soft LVQ (RSLVQ). These learning algorithms, however, are restricted to Euclidean vectors. Thus they are unsuitable for complex or heterogeneous data sets where input dimensions have different relevance or a high dimensionality yields to accumulated noise which can disrupt the classifications. Although this problem can be partially avoided by appropriate metric learning or by kernel variants, however, if data are inherently non-Euclidean, these techniques can not be applied. In addition, in modern applications, data are often addressed using dedicated non-Euclidean dissimilarities such as dynamic time warping for time series, alignments for symbolic strings, the compression distance to compare sequences based on an information theoretic ground, and similar. These settings do not allow an Euclidean representation of data at all, rather data are given implicitly in terms of pairwise dissimilarities or relations.

In the contribution, we propose extensions of GLVQ and RSLVQ, which can directly deal with relational data sets which are characterized in terms of a symmetric dissimilarity matrix only. The optimization can take place using gradient techniques. We test these techniques on several biomedical benchmark data sets, and the results are comparable to SVM while providing prototype based presentation.

3.27 Agents Learning a Complex Task/ Dispersive PSO

Jort van Mourik (Aston University – Birmingham, GB)

Joint work of van Mourik, Jort; Goldingay, Harry
Main reference H. Goldingay, J. van Mourik, "The effect of load on agent-based algorithms for distributed task allocation," Information Sciences, Elsevier.
URL http://dx.doi.org/10.1016/j.ins.2011.06.011

We present an overview of recent developments concerning various algorithms based on autonomous agents learning a complex optimisation task. We show that relatively simple algorithms based on autonomous agents can be competitive with the best centralised optimisation algorithms for which communication costs may soon become prohibitive for large systems. We propose a hybrid algorithm that combines the best features of both threshold- and market- based algorithms, and by introduction of a simple form of memory we obtain 98.5% of the theoretical efficiency limit of the optimal centralised algorithm, while keeping excellent scalability. As various algorithms are compared, each of which has a set of parameters that need tuning for fair comparison, the need for a good general purpose optimisation method arises. We propose a version of particle swarm optimisation (PSO) that avoids early convergence: Dispersive PSO. We show that this form of PSO generally outperforms existing ones in cases where early convergence (to a local optimum) is an issue.

3.28 Machine Learning for Data Visualization

Laurens van der Maaten (TU Delft, NL)

The talk gives an (incomplete) overview of machine-learning techniques that can be used for the visualization of high-dimensional data. In particular, it focuses on two types of dimensionality reduction techniques: (1) generative models that reduce dimensionality by performing maximum-likelihood learning in a (non)linear Gaussian model and (2) manifold learners that reduce dimensionality by preserving local properties of the data manifold. The last part of the talk focuses on some problems that arise specifically in visualization settings; for instance, with the question of how to visualize non-metric similarities or contingency tables.

4 Preliminary follow up publications resulting from the seminar

4.1 Supervised learning of short and high-dimensional temporal sequences for life science measurements

Frank-Michael Schleif (Universität Bielefeld, DE)

Joint work of F.-M.Schleif, A. Gisbrecht; B. Hammer
Main reference F.-M. Schlief, A. Gisbrecht, B. Hammer, "Supervised learning of short and high-dimensional temporal sequences for life science measurements," Dagstuhl Preprint Archive, arXiv:1110.2416v1 [cs.LG].
URL http://arxiv.org/abs/1110.2416v1

The analysis of physiological processes over time are often given by spectrometric or gene expression profiles over time with only few time points but a large number of measured variables. The analysis of such temporal sequences is challenging and only few methods have been proposed. The information can be encoded time independent, by means of classical expression differences for a single time point or in expression profiles over time. Available methods are limited to unsupervised and semi-supervised settings. The predictive variables can be identified only by means of wrapper or post-processing techniques. This is complicated due to the small number of samples for such studies. Here, we present a supervised learning approach, termed Supervised Topographic Mapping Through Time (SGTM-TT). It learns a supervised mapping of the temporal sequences onto a low dimensional grid. We utilize a hidden markov model (HMM) to account for the time domain and relevance learning to identify the relevant feature dimensions most predictive over time. The learned mapping can be used to visualize the temporal sequences and to predict the class of a new sequence. The relevance learning permits the identification of discriminating masses or gen expressions and prunes dimensions which are unnecessary for the classification task or encode mainly noise. In this way we obtain a very efficient learning system for temporal sequences. The results indicate that using simultaneous supervised learning and metric adaptation significantly improves the prediction accuracy for synthetically and real life data in comparison to the standard techniques. The discriminating features, identified by relevance learning, compare favorably with the results of alternative methods. Our method permits the visualization of the data on a low dimensional grid, highlighting the observed temporal structure.

4.2 PAC learnability versus VC dimension: a footnote to a basic result of statistical learning

Vladimir Pestov (University of Ottawa, CA)

Main reference V. Pestov, "PAC learnability versus VC dimension: a footnote to a basic result of statistical learning," arXiv:1104.2097v1 [cs.LG].
URL http://arxiv.org/abs/1104.2097v1

A fundamental result of statistical learning theory states that a concept class is PAC learnable if and only if it is a uniform Glivenko-Cantelli class if and only if the VC dimension of the class is finite. However, the theorem is only valid under special assumptions of measurability of the class, in which case the PAC learnability even becomes consistent. Otherwise, there is a classical example, constructed under the Continuum Hypothesis by Dudley and Durst and further adapted by Blumer, Ehrenfeucht, Haussler, and Warmuth, of a concept class of VC dimension one which is neither uniform Glivenko-Cantelli nor consistently PAC learnable.

We show that, rather surprisingly, under an additional set-theoretic hypothesis which is much milder than the Continuum Hypothesis (Martin's Axiom), PAC learnability is equivalent to finite VC dimension for every concept class.

Comments: Revised submission to IJCNN 2011.

4.3 How to Evaluate Dimensionality Reduction?

Bassam Mokbel

Joint work of Wouter Lueks, Michael Biehl, Barbara Hammer
Main reference W. Lueks, B. Mokbel, M. Biehl, B. Hammer, "How to Evaluate Dimensionality Reduction? – Improving the Co-ranking Matrix," Dagstuhl Preprint Archive, arXiv:1110.3917v1 [cs.LG].
URL http://arxiv.org/abs/1110.3917v1

The growing number of dimensionality reduction methods available for data visualization has recently inspired the development of quality assessment measures, in order to evaluate the resulting low-dimensional representation independently from a methods' inherent criteria. Several (existing) quality measures can be (re)formulated based on the so-called co-ranking matrix, which subsumes all rank errors (i.e. differences between the ranking of distances from every point to all others, comparing the low-dimensional representation to the original data). The measures are often based on the partitioning of the co-ranking matrix into 4 submatrices, divided at the K-th row and column, calculating a weighted combination of the sums of each submatrix. Hence, the evaluation process typically involves plotting a graph over several (or even all possible) settings of the parameter K. Considering simple artificial examples, we argue that this parameter controls two notions at once, that need not necessarily be combined, and that the rectangular shape of submatrices is disadvantageous for an intuitive interpretation of the parameter. We debate that quality measures, as general and flexible evaluation tools, should have parameters with a direct and intuitive interpretation as to which specific error types are tolerated or penalized. Therefore, we propose to replace K with two parameters to control these notions separately, and introduce a differently shaped weighting on the co-ranking matrix. The two new parameters can then directly be interpreted as a threshold up to which rank errors are tolerated, and a threshold up to which the rank-distances are significant for the evaluation. Moreover, we propose a color representation of local quality to visually support the evaluation process for a given mapping, where every point in the mapping is colored according to its local contribution to the overall quality.

4.4 About Generalization of the Conn-Index for Fuzzy Clustering Validation

Thomas Villmann (Computational Intelligence Group, University of Applied Sciences Mittweida, DE)

Joint work of T. Geweniger, M.ästner, M. Lange, and T. Villmann
Main reference T. Geweniger, M. Kästner, M. Lange, and T. Villmann, "Derivation of a Generalized Conn-Index for Fuzzy Clustering Validation," Machine Learning Reports, Report 07/2011.
URL http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_07_2011

Clustering and cluster validation strongly depends on the underlying model, the used dissimilarity measure, complexity constraints and other limiting options. In context of very high dimensional data and large data sets preprocessing and compression of the data by vector quantization is one possibility to deal with this complexity. Subsequent clustering of the compressed data should take into account this information as well as cluster validation. One approach in this direction and presented at the seminar is the so-called Conn-Index invented by *Erzsébet Merényi* and *Kadim Taşdemir* [11]. In this approach, for the evaluation of the clusters consisting of vector quantization prototypes, the topological structure information between the prototype vectors acquired during the vector quantization learning is used to assess the quality of the cluster solution. This information is available by the Delaunay-graph with respect to the Voronoi tessellation of the data space according to the prototypes.

We discussed in a participant group (E. Merényi, T. Geweniger, M. Kästner, K. Taşdemir, and T. Villmann), how an extension of this approach could be designed such that fuzzy clustering also would be covered. Fuzzy vector quantization is mainly influenced by the fuzzy c-means algorithm for probabilistic fuzzy assignments [1, 5], its probabilistic counterpart [7, 8], and variants integrating neighborhood cooperativeness for better stability and convergence [3, 2, 4, 9, 10, 12, 13]. In consequence of these discussion we agreed that the topological structure between the prototypes in fuzzy vector quantization is implicitly contained in the fuzzy assignments, and, therefore, could be used to extend the Conn-Index for those vector quantization models.

As a preliminary result we can present a first publication, where we stated these thoughts more precisely formulating the underlying theoretical concepts of the new fuzzy Conn-index, see [6].

References

1 J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum, New York, 1981.

2 J. C. Bezdek and N. R. Pal. A note on self-organizing semantic maps. *IEEE Transactions on Neural Networks*, 6(5):1029–1036, 1995.

3 J. C. Bezdek and N. R. Pal. Two soft relatives of learning vector quantization. *Neural Networks*, 8(5):729–743, 1995.

4 J. C. Bezdek, E. C. K. Tsao, and N. R. Pal. Fuzzy Kohonen clustering networks. In *Proc. IEEE International Conference on Fuzzy Systems*, pages 1035–1043, Piscataway, NJ, 1992. IEEE Service Center.

5 J. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

6 T. Geweniger, M. Kästner, M. Lange, and T. Villmann. Derivation of a generalized Conn-index for fuzzy clustering validation. *Machine Learning Reports*, 5(MLR-07-2011):1–12, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/ ˜ fschleif/mlr/mlr_07_2011.pdf.

7 R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(4):98–110, 1993.

8 N. Pal, K. Pal, J. Keller, and J. Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530, 2005.

9 N. R. Pal, J. C. Bezdek, and E. C. K. Tsao. Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks*, 4(4):549–557, 1993.

10 N. R. Pal, J. C. Bezdek, and E. C. K. Tsao. Errata to Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks*, 6(2):521–521, March 1995.

11 K. Taşdemir and E. Merényi. A validity index for prototype-based clustering of data sets with complex cluster structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(4):1039 − −1053, 2011.

12 E. Tsao, J. Bezdek, and N. Pal. Fuzzy Kohonen clustering networks. *Pattern Recognition*, 27(5):757–764, 1994.

13 T. Villmann, T. Geweniger, M. Kästner, and M. Lange. Theory of fuzzy neural gas for unsupervised vector quantization. *Machine Learning Reports*, 5(MLR-06-2011):27–46, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_06_2011.pdf.

Participants

- Gyan Bhanot
Rutgers Univ. – Piscataway, US
- Michael Biehl
University of Groningen, NL
- Kerstin Bunte
University of Groningen, NL
- Gert-Jan de Vries
Philips Research Lab. – Eindhoven, NL
- Klaus Dohmen
Hochschule Mittweida, DE
- Richard Farkas
Universität Stuttgart, DE
- Tina Geweniger
University of Groningen, NL
- Andrej Gisbrecht
Universität Bielefeld, DE
- Sven Haase
Hochschule Mittweida, DE
- Barbara Hammer
Universität Bielefeld, DE
- Marika Kästner
Hochschule Mittweida, DE
- Arto Klami
Aalto University, FI
- Neil D. Lawrence
Sheffield University, GB
- John A. Lee
University of Louvain, BE
- Marco Loog
TU Delft, NL
- Thomas Martinetz
Universität Lübeck, DE
- Erzsébet Merényi
Rice University, US
- Bassam Mokbel
Universität Bielefeld, DE
- Ernest Mwebaze
Makerere Univ. – Kampala, SAF
- Oliver Obst
CSIRO ICT Centre – Marsfield, AU
- Vladimir Pestov
University of Ottawa, CA
- John Quinn
Makerere Univ. – Kampala, SAF
- Fabrice Rossi
Télécom Paris Tech, FR
- Frank-Michael Schleif
Universität Bielefeld, DE
- Petra Schneider
University of Birmingham, GB
- Udo Seiffert
Fraunhofer IFF Magdeburg, DE
- Diego Sona
Fondazione Bruno Kessler – Trento, IT
- Marc Strickert
Universität Siegen, DE
- Kadim Tasdemir
EC Joint Research Centre – Ispra, IT
- David M. J. Tax
TU Delft, NL
- Peter Tino
University of Birmingham, GB
- Laurens van der Maaten
TU Delft, NL
- Jort van Mourik
Aston Univ. – Birmingham, GB
- Marina Vanucci
Rice University – Houston, US
- Michel Verleysen
UC Louvain-la-Neuve, BE
- Thomas Villmann
Hochschule Mittweida, DE
- Kilian Weinberger
Washington University, US
- Xibin Zhu
Universität Bielefeld, DE
- Dietlind Zühlke
Fraunhofer Institut FIT – St. Augustin, DE