Report from Dagstuhl Seminar 12081

Information Visualization, Visual Data Mining and Machine Learning

Edited by

Daniel A. Keim[1], Fabrice Rossi[2], Thomas Seidl[3], Michel Verleysen[4], and Stefan Wrobel[5]

1 **Universität Konstanz, DE,** keim@uni-konstanz.de
2 **SAMM, Université Paris 1, FR,** Fabrice.Rossi@univ-paris1.fr
3 **RWTH Aachen, DE,** seidl@informatik.rwth-aachen.de
4 **Université Catholique de Louvain, BE,** michel.verleysen@uclouvain.be
5 **Fraunhofer IAIS - St. Augustin, DE and University of Bonn, DE,** stefan.wrobel@iais.fraunhofer.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 12081 “Information Visualization, Visual Data Mining and Machine Learning”. The aim of the seminar was to tighten the links between the information visualisation community and the machine learning community in order to explore how each field can benefit from the other and how to go beyond current hybridization successes.

Seminar *19.–24. February, 2012* – http://www.dagstuhl.de/12081
1998 ACM Subject Classification H.5 Information interfaces and presentations, I.2.6 Learning, H.2.8 Database Applications, H.3.3 Information Search and Retrieval, I.5.3 Clustering.
Keywords and phrases Information visualization, visual data mining, machine learning, nonlinear dimensionality reduction, exploratory data analysis.
Digital Object Identifier 10.4230/DagRep.1.1.1

1 Executive Summary

Daniel A. Keim, Fabrice Rossi, Thomas Seidl, Michel Verleysen, and Stefan Wrobel

Information visualization and visual data mining leverage the human visual system to provide insight and understanding of unorganized data. Visualizing data in a way that is appropriate for the user’s needs proves essential in a number of situations: getting insights about data before a further more quantitative analysis, presenting data to a user through well-chosen table, graph or other structured representations, relying on the cognitive skills of humans to show them extended information in a compact way, etc.

Machine learning enables computers to automatically discover complex patterns in data and, when examples of such patterns are available, to learn automatically from the examples how to recognize occurrences of those patterns in new data. Machine learning has proven itself quite successful in day to day tasks such as SPAM filtering and optical character recognition.

Both research fields share a focus on data and information, and it might seem at first that the main difference between the two fields is the predominance of visual representations of the

DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

data in information visualization compared to its relatively low presence in machine learning. However, it should be noted that visual representations are used in a quite systematic way in machine learning, for instance to summarize predictive performances, i.e., whether a given system is performing well in detecting some pattern. This can be traced back to a long tradition of statistical graphics for instance. Dimensionality reduction is also a major topic in machine learning: one aims here at describing as accurately as possible some data with a small number of variables rather than with their original possibly numerous variables. Principal component analysis is the simplest and most well known example of such a method. In the extreme case where one uses only two or three variables, dimensionality reduction is a form of information visualization as the new variables can be used to directly display the original data.

The main difference between both fields is the role of the user in the data exploration and modeling. The ultimate goal of machine learning is somehow to get rid of the user: everything should be completely automated and done by a computer. While the user could still play a role by, e.g., choosing the data description or the type of algorithm to use, his/her influence should be limited to a strict minimum. In information visualization, a quite opposite point of view is put forward as visual representations are designed to be leveraged by a human to extract knowledge from the data. Patterns are discovered by the user, models are adjusted to the data under user steering, etc.

This major difference in philosophy probably explains why machine learning and information visualization communities have remained relatively disconnected. Both research fields are mature and well structured around major conferences and journals. There is also a strong tradition of Dagstuhl seminars about both topics. Yet, despite some well known success, collaboration has been scarce among researchers coming from the two fields. Some success stories are the use of state-of-the-art results from one field in the other. For instance, Kohonen's Self Organizing Map, a well known dimensionality reduction technique, has been successful partly because of its visualization capabilities which were inspired by information visualization results. In the opposite direction, information visualization techniques often use classical methods from machine learning, for instance, clustering or multidimensional scaling.

The seminar was organized in this context with the specific goal of bringing together researchers from both communities in order to tighten the loose links between them. To limit the risk of misunderstandings induced by the different backgrounds of researchers from the two communities, the seminar started with introductory talks about both domains. It was then mainly organized as a series of thematic talks with a significant portion of the time dedicated to questions and discussions. After the first two days of meeting, understanding between both communities reached a sufficient level to organize, in addition to the plenary talks, working group focusing on specific issues.

Several research topics emerged from the initial discussions and lead to the creation of the working groups. The subject that raised probably the largest number of questions and discussions is Evaluation. It is not very surprising as differences between the communities about evaluation (or quality assessment) might be considered as the concrete technical manifestation of cultural and philosophical differences between them. Indeed, in machine learning, automatic methods are mostly designed according to the following general principle: Given a quality measure for a possible solution of the problem under study, one devises an algorithm that searches the solution space efficiently for the optimal solution with respect to this measure. For instance, in SPAM filtering a possible quality measure is the classification accuracy of the filter: it has to sort unsolicited bulk messages correctly into the SPAM class and all other emails in the HAM class. In a simple setting, the best filter could be considered as the one with the smallest number of errors. However, counting only the number of errors is usually too naive, and better quality measures have to be used, such as the area under the ROC curve: the Receiver Operating Characteristic curve shows the dependency between the true positive rate (the percentage of unsolicited bulk messages classified as SPAM) and the false positive rate (the percentage of correct emails classified as SPAM).

In information visualization, evaluation cannot rely only on mathematical quality measures as the user is always part of the story. A successful visualization is a solution, with which the user is able to perform better, in a general sense, compared to existing solutions. As in machine learning, a method is therefore evaluated according to some goal and with some quality metric, but the evaluation process and the quality metrics have to take the user into account. For instance, one display can be used to help the user assess the correlation between variables. Then, a quality metric might be the time needed to find a pair of highly correlated variables, or the time needed to decide that there is no such pair. Another metric might be the percentage of accurate decisions about the correlation of some pairs of variables. In general, a visualization system can be evaluated with respect to numerous tasks and according to various metrics. This should be done in a controlled environment and with different users, to limit the influence of interpersonal variations.

Among the discussions between members of the two communities about evaluation, questions were raised about the so-called unsupervised problems in machine learning. These problems, such as clustering or dimensionality reduction, are ill-posed in a machine learning sense: there is no unquestionable quality metric associated to e.g. clustering but rather a large number of such metrics. Some of those metrics lead to very difficult optimization problems (from a computational point of view) that are addressed via approximate heuristic solutions. In the end, machine learning has produced dozens of clustering methods and dimensionality reduction methods, and evaluations with respect to user needs remain an open problem. An important outcome of the seminar was to reposition this problem in the global picture of collaboration between information visualization and machine learning. For instance, if many quality measures are possible, one way to compare them would be to measure their link to user performances in different tasks. If several methods seem to perform equally well in a machine learning sense, then the user feedback could help to indentify the «best» method. It was also noted that many methods that are studied in machine learning and linked to information visualization, in particular dimensionality reduction and embedding techniques, would benefit from more interaction between the communities. At minimum, state-of-the-art methods from machine learning should be known by information visualization researchers and state-of-the-art visualization techniques should be deployed by machine learning researchers.

Another topic discussed thoroughly at the seminar was the visualization of specific types of objects. Relational data were discussed, for instance, as a general model for heterogeneous complex data as stored in a relational database. Graph visualization techniques provide a possible starting point, but it is clear that for large databases, summarization is needed, which brought back the discussion of the ill defined clustering problem mentioned above. Among complex objects, models obtained by a machine learning algorithms were also considered, in particular as good candidates for interactive visualizations. Decision trees give a good example of such objects: Given a proper visualization of the current tree, of some possible simplified or more complex versions and of the effect of the tree(s) on some dataset, an expert user can adapt the tree to his/her specific goals that are not directly expressible in a quality criterion. The extreme case of visualizing the dynamic evolution of a self learning process was discussed as a prototype of complex objects representation: The system is evolving through time, it learns decision rules, and it evolves using complex (and evolving) decision tables.

Finally, it became clear that a large effort is still needed at the algorithmic and software levels. First, fast machine learning techniques are needed that can be embedded in interactive visualization systems. Second, there is the need for a standard software environment that can be used in both communities. The unavailability of such a system hurts research to some extent as some active system environments in one field do not include even basic facilities from the other. One typical example is the R statistical environment with which a large part of machine learning research is conducted and whose interactive visualization capabilities are limited, in particular in comparison to the state-of-the-art static visualization possibilities. One possible solution foreseen at the seminar was the development of some dynamic data sharing standard that can be implemented in several software environments, allowing fast communication between those environments and facilitating software reuse.

Judging by the liveliness of the discussions and the number of joint research projects proposed at the end of the seminar, this meeting between the machine learning and the information visualization communities was more than needed. The flexible format of the Dagstuhl seminars is perfectly adapted to this type of meeting and the only frustration perceivable at the end of the week was that it had indeed reached its end. It was clear that researchers from the two communities were starting to understand each other and were eager to share more thoughts and actually start working on joint projects. This calls for further seminars ...

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Graph visualization methods and data mining: results, evaluation, and future directions

Daniel Archambault (University College Dublin, IE)

Graph visualization and data mining methods have many areas of common interest.

In the introductory talk for this session, I will cover some of my recent results on graph visualization applicable to this topic, outline methods of visualization research, and identify some possible areas of future collaboration.

3.2 Steerable Large Scale Data Analytics

Daniel Archambault (University College Dublin, IE)

In this short talk, I cover some ideas on steerable data analytics. In this area, I think that we should strive to strengthen the coupling between data mining or clustering processes and visualization in order to enable real time analysis. I give potential ways to achieve this goal with possible applications to the area of social media analysis and community finding.

3.3 Multivariate data exploration with CheckViz and ProxiViz

Michael Aupetit (Commissariat a l'Energie Atomique - Gif-sur-Yvette, FR)

Joint work of Aupetit, Michael; Lespinats, Sylvain;
Main reference Sylvain Lespinats, and Michaël Aupetit, CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings. Computer Graphics Forum 30(1):113-125, 2011.
URL http://dx.doi.org/10.1111/j.1467-8659.2010.01835.x

Embedding techniques are used for multivariate data analysis. They provide a planar set of points whose relative distances estimates the original similarities.

We argue that this set of points alone is not enough to make sense out of it. We present CheckViz [2] and ProxiViz [1] as two ways to make the set of points interpretable by the user. CheckViz overload distortions straight into the map, it can be used as a sanity check and also provides inference rule which help to recover the original data topology. ProxiViz overload the true original similarity measure between a selected point and each of the other points which makes possible to reconstruct the original data structure. The embeddings appear not to be an end, but just a mean to display a complementary information which make them usable and useful for multivariate data exploration.

References

1 Michaël Aupetit, *Visualizing distortions and recovering topology in continuous projection techniques.* Neurocomputing 70(7-9):1304-1330, March 2007.

2 Sylvain Lespinats, and Michaël Aupetit, *CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings.* Computer Graphics Forum 30(1):113-125, 2011.

3.4 Matrix relevance learning and visualization of labeled data sets

Michael Biehl (University of Groningen, NL)

Joint work of Biehl, Michael; Bunte, Kerstin; Hammer, Barbara; Schneider, Petra; Villmann, Thomas;

A brief introduction is given to Learning Vector Quantization (LVQ) as an intuitive, flexible, and very powerful prototype-based classifier.

The focus is on the recent extension of LVQ by Matrix Relevance Learning. In this scheme, one or several matrices of adaptive relevances are employed to parameterize a distance measure.

Matrix Relevance Learning makes use of a low-dimensional linear or locally linear representation of the data set, internally. This fact can be exploited for the discriminative visualization of labelled data sets.

In terms of a few application examples from the life sciences it is argued that these visualizations facilitate valuable insight into the nature of the problems.

Possible routes to extend the schemes to explicitly non-linear visualizations are briefly discussed. This leads to the question what the goal of visualizing labeled data should be.

The following references may serve as a starting point to get acquainted with Matrix Relevance Learning in the context of visualization.

References

1 P. Schneider, M. Biehl, and B. Hammer. *Adaptive relevance matrices in Learning Vector Quantization.*
Neural Computation 21(12): 3532-3561, 2009

2 K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl. *Limited rank matrix learning, discriminative dimension reduction, and visualization*. Neural Networks 26: 159-173, 2012

3.5 Supervised dimension reduction - A brief history

Kerstin Bunte (Universität Bielefeld, DE)

Joint work of Bunte, Kerstin; Biehl, Michael; Hammer, Barbara

Due to improved sensor technology, dedicated data formats and rapidly increasing digitalization capabilities the amount of electronic data increases dramatically since decades. As a consequence the manual inspection data sets often becomes infeasible. In recent years, many powerful non-linear dimension reduction techniques have been developed which provide a visualization of complex data sets. Using prior knowledge, e.g. in form of supervision might provide more informative mappings dependent on the actual data set.

3.6 Overview of Visual Inference

Di Cook (Iowa State University, US)

Main reference Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, Hadley Wickham. Statistical Inference for Exploratory Data Analysis and Model Diagnostics. Royal Society Philosophical Transactions A, vol. 367, no. 1906, pp. 4361–4383, 2009.
URL http://dx.doi.org/10.1098/rsta.2009.0120

Implicitly detection of patterns in a plot of data is a rejection of some null hypothesis. What patterns might we see in the plot if the data was sampled in a manner consistent with the null hypothesis? This research area provides methods for assessing whether what we see in plots is "real", and obtaining levels of significance for findings based on visualization. Two protocols are used, a lineup and a rorschach. In the lineup, the plot of real data is embedded in a field of plots of data generated in a manner consistent with the relevant null hypothesis. In a rorschach, all plots are null plots, and the approach is a way to examine how much variability can occur purely by chance.

3.7 Eye-tracking Experiments for Visual Inference

Di Cook (Iowa State University, US)

Visual inference provides methods for assessing whether what we see in plots is real. The primary method is a lineup, where a plot of the actual data is embedded in a field of plots of data generated in a manner consistent with the null hypothesis. For example, to assess the relationship between two variables with a scatterplot, the null plots may show the same data, with one of the variables having its values permuted, thus breaking any real association between the avriables. If the observer picks the actual data plot from the lineup it lends significance to the conclusion of a real relationship between the two variables. Following a series of Amazon Turk experiments where the lineup protocol was evaluated under controlled simulated data experimental conditions, we selected a handful of lineups for detailed assessment. Here subjects were recorded with an eye trackers to examine (1) how long they looked at their selection, (2) which plots caught the subjects attention and (3) how subjects scanned the lineups to make their selections.

3.8 Future Analysis Environments

Jean-Daniel Fekete (Université Paris Sud - Orsay, FR)

What is the future of Analysis environments allowing machine learning and visualization to interoperate seamlessly? Should we design a new system that will solve all the problems, reuse already existing systems or in-between? These slides summarize a possible way to address the issue that might address the problem is a simple enough way.

3.9 Psychology of Visual Analytics

Brian D. Fisher (Simon Fraser University - Surrey, CA)

Main reference Fisher, B., Green, T.M., Arias-Hernández, R. (2011) Visual Analytics as a Translational Cognitive Science, Topics in Cognitive Science 3,3 609-625.
URL http://dx.doi.org/10.1111/j.1756-8765.2011.01148.x

This talk explores the larger implications of visual analytics – the science of analytical reasoning facilitated by interactive visual interfaces for cognitive science and informatics. The visual analytics approach emphasizes the design of technologies to support the ability of trained human analysts to understand situations, make decisions, generate plans, and put them into action. The resulting visual information systems succeed when they enable analysts to more effectively work with complex "big data" from sensors, archives, computational and mathematical models, alone and in collaboration with other analysts.

My laboratory begins by building field study methods that characterize human and computational cognitive capabilities as they are used for decision-making in specific situations in flight safety, public health, and emergency management analysis. These field studies generate research questions and experimental protocols that are used to investigate human-computer cognitive systems in the laboratory. My talk will briefly discuss our "pair analytics" methods derived from H. Clark's Joint Activity Theory and two laboratory studies of perceptual cognition in display environments similar to those proposed for air traffic control and collaborative aircraft CAD.

3.10 BCI-based Evaluation in Information Visualization

Hans Hagen (TU Kaiserslautern, DE)

Joint work of Hagen, Hans; Ebert, Achim; Cernea, Daniel;
Main reference Daniel Cernea, Peter-Scott Olech, Achim Ebert, and Andreas Kerren. Measuring Subjectivity “Supporting Evaluations with the Emotiv EPOC Neuroheadset. Kuenstliche Intelligenz” KI Journal, Special Issue on Human-Computer Interaction, Volume 26, Number 2 (2012), 177-182.
URL http://dx.doi.org/10.1007/s13218-011-0165-0

Evaluations have been the key factor for validating different visualization and interaction approaches. But while experts agree on their importance, the evaluation techniques currently used in Information Visualization focus mostly on objective measurements like performance and efficiency, and only rarely investigate subjective factors (states of mind and emotions that the users experience).

As the ideal evaluation should be non-intrusive and executed in real-time, many researchers turn to novel brain-computer interfaces (BCI) for directly investigating the users' affective and mental states. While current portable BCI systems are employed overwhelmingly in control tasks (e.g. moving a robotic Arm), many of them have proven useful in supporting subjectivity measurements and, thus, evaluations in real-time.

But what would an ideal BCI system detect and how would it process it in order to support the evaluation of Information Visualization approaches? Could a framework specifically designed for InfoVis evaluation with BCI systems enable researchers to obtain the answers they seek? These are a couple of specific topics that need to be addressed when looking at the potential of BCI systems as an alternative evaluation method for Information Visualization techniques and systems.

3.11 Including prior knowledge into data visualization

Barbara Hammer (Universität Bielefeld, DE)

In this presentation, the question of how data visualization and dimensionality reduction are linked to prior knowledge will be investigated. First, it will be motivated, that visualization and prior knowledge are closely connected.

Afterwards, technical possibilities how to integrate different kind of prior knowledge into dimensionality reduction will be discussed. Four major principles will be identified and demonstrated by examples: (i) change of the prior in a Bayesian model. For a cost-function based techniques, possibilities are given by a (ii) change of the data representation or metric, (iii) change of the cost function used for training, (iv) change or bias of the mapping of data to low dimensions.

3.12 Automated Methods in Information Visualization

Helwig Hauser (University of Bergen, NO)

Visualization and Machine Learning have related goals in terms of helping analysts to understand characteristic aspects of data. While visualization aims at involving the user through interactive depictions of data, machine learning is generally represented by automatic methods that yield optimal results with respect to certain initially specified tasks. Not at the least within the research direction of visual analytics it seems promising to think about opportunities to integrate both methodologies in order to exploit the strengths of both sides. Up to now, examples of integration very often encompass the visualization of results from automatic methods as well as attempts to make originally automated methods partially interactive. A vision for the future would be to integrate interactive and automatic methods in order to solve problems. A possible realization could be an iterative process where the one or other approach is chosen on demand at each step.

3.13 Distance concentration and detection of meaningless distances

Ata Kaban (University of Birmingham, GB)

Main reference A Kaban. Non-parametric Detection of Meaningless Distances in High-Dimensional Data. Statistics and Computing. 22(1): 375-385.
URL http://dx.doi.org/10.1007/s11222-011-9229-0

Distance concentration is a counter-intuitive aspect of the curse of dimensionality, the phenomenon that in certain conditions the contrast between the nearest and the farthest neighbouring points vanishes as the data dimension increases. This makes distances meaningless, exponentially slows down data retrieval, and risks to compromise our ability to extract meaningful information from high dimensional data sets. First, we show that the known sufficient conditions are also the necessary conditions of distance concentration in the limit of infinite dimensions. We then quantify the phenomenon more precisely, for possibly high but finite dimensional settings in a distribution-free manner, by bounding the tails of the probability that distances become meaningless. We show how this can be turned into a statistical test to assess the concentration of a given distance function in some unknown data distribution solely on the basis of an available data sample from it. This can be used to test and detect problematic cases more rigorously than it has been possible previously, and we demonstrate the working of this approach on both synthetic data and ten real-world data sets from different domains.

References

1 A Kaban. *Non-parametric Detection of Meaningless Distances in High-Dimensional Data*. Statistics and Computing. 22(1): 375-385.

3.14 Visual Analysis of Multi-faceted Scientific Data: a Survey

Johannes Kehrer (VRVis - Wien, AT)

Joint work of Kehrer, Johannes; Hauser, Helwig;
Main reference Johannes Kehrer and Helwig Hauser, Visualization and Visual Analysis of Multi-faceted Scientific Data: A Survey, IEEE Trans. on Visualization and Computer Graphics, accepted for publication.
URL http://dx.doi.org/10.1109/TVCG.2012.110

Interactive visual analysis plays an important role in studying different kinds of scientific data (e.g., spatial, temporal and/or multi-variate data). The talk is based on a thorough literature review, which investigates to which degree methods for 1) visual representation, 2) user interaction and 3) computational analysis are combined in such an analysis. A task-based categorization of approaches is proposed and different options for the visual analysis are discussed. This leads to conclusions with respect to promising research directions, for instance, to pursue new solutions that combine supervised machine learning with interactive feature specification via brushing.

3.15 Towards Visual Analytics

Daniel A. Keim (Universität Konstanz, DE)

Joint work of Keim, Daniel A.; Kohlhammer, Joern; Geoffrey, Ellis; Mansmann, Florian
Main reference D. Keim, J. Kohlhammer, G. Ellis, F. Mansmann (editors): Mastering the Information Age - Solving Problems with Visual Analytics, Eurographics, 2010
URL http://www.vismaster.eu/book/

Many of the grand challenges require not only automatic methods, but also exploration to find appropriate solutions. Visual Analytics as the tight integration of visual and automatic data analysis methods for information exploration and scalable decision support aims at integrating machine capabilities (e.g., data storage, numerical computation or search) with human capabilities (such as perception, creativity and general knowledge). Besides giving an introduction to Visual Analytics and information visualization, this talk describes common evaluation approaches and outlines the relation between visualization and machine learning.

3.16 Visualization of Network Centralities

Andreas Kerren (Linnaeus University - Växjö, SE)

Joint work of Kerren, Andreas; Köstinger, Harald; Zimmer, Björn;
Main reference Andreas Kerren, Harald Köstinger, and Björn Zimmer. ViNCent "Visualization of Network Centralities." In Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP '12), pages 703-712, Rome, Italy, 2012. INSTICC.

The use of network centralities in the field of network analysis plays an important role when the relative importance of nodes within the network topology should be rated. A single network can easily be represented by the use of standard graph drawing algorithms, but not only the exploration of one centrality might be important: the comparison of two or more of them is often crucial for a better understanding. When visualizing the comparison of several network centralities, we are facing new problems of how to show them in a meaningful way. For instance, we want to be able to track all the changes of centralities in the networks as well as to display the single networks as best as possible. In the life sciences, centrality measures help scientists to understand the underlying biological processes and have been successfully applied to different biological networks. The aim of this talk was to briefly present a system for the interactive visualization of biochemical networks and its centralities. Researchers can focus on the exploration of the centrality values including the network structure without dealing with visual clutter or occlusions of nodes. Simultaneously, filtering based on statistical data concerning the network elements and centrality values supports this.

3.17 Embedding from high- to low-dimensional spaces; how can we cope with the phenomenon of norm concentration?

John A. Lee (Université Catholique de Louvain, BE)

Joint work of Lee, John A.; Verleysen, M.;
Main reference John A. Lee, Michel Verleysen, Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants, Procedia Computer Science Volume 4, 2011, Pages 538–547.
URL http://dx.doi.org/10.1016/j.procs.2011.04.056

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 paper also proposes a generalized definition of shift-invariant similarities that extend the applicability of SNE to noisy data.

3.18 Visual Analytics of Sparse Data

Marcus A. Magnor (TU Braunschweig, DE)

High-dimensional data will always constitute only sparse representations of inter-dimensional information. As a result of large voids in n-D space, even without taking noise and erroneous data into account, putative inter-dimensional relations may only be halluscinated, by humans as well as by algorithms. In contrast, suitable interpolation on the data level, guided by high-level knowledge of the data and dimensional meaning, may be able to plausibly fill the voids and to fortify subsequent interactive and automatic analysis results.

3.19 Exploration through Enrichment

Florian Mansmann (Universität Konstanz, DE)

Joint work of Mansmann, Florian; Spretke, David; Janetzko, Halldor; Bak, Peter
Main reference D. Spretke, H. Janetzko, F. Mansmann, P. Bak, B. Kranstauber, S. Davidson and M. Mueller. Exploration through Enrichment: A Visual Analytics Approach for Animal Movement.

Proceedings of the 19th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 421–424, 2011.
URL http://dx.doi.org/10.1145/2093973.2094038

In many visualization scenarios, visualizing and exploring data raises hypotheses that cannot be answered with the current data. Therefore, very often an enrichment phase is needed to enhance the exploration process. In this talk, I showed two prototypes, namely ClockView in which network time series can be filtered through user-defined patterns and the Animal Ecology Explorer in which bird movement can be interactively refined through machine learning methods such as clustering and classification.

References

1 C. Kintzel, J. Fuchs and F. Mansmann. *Monitoring Large IP Spaces with ClockView*. Proc. of Int. Symp. on Visualization for Cyber Security (VizSec), 2011.

2 D. Spretke, H. Janetzko, F. Mansmann, P. Bak, B. Kranstauber, S. Davidson and M. Mueller. *Exploration through Enrichment: A Visual Analytics Approach for Animal Movement*. Proc. of the 19th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 421–424, 2011.

3.20 Quality metrics for InfoVis

Florian Mansmann (Universität Konstanz, DE)

Joint work of Bertini, Enrico; Tatu, Andrada; Keim, Daniel A.
Main reference E. Bertini, A. Tatu and D. A. Keim. Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. IEEE Symposium on Information Visualization (InfoVis), 2011.
URL http://dx.doi.org/10.1109/TVCG.2011.229

Quality metrics are a recent trend in the information visualization community.

The basic idea is that the quality of a visualization with respect to the loaded data can be calculated and based on this assessment the good or optimal parameter configurations for visualizations can be found.

References

1 E. Bertini, A. Tatu and D. A. Keim. *Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization*. IEEE Transactions on Visualization and Computer Graphics (TVCG), vol. 17, no. 12, pp. 2203-2212, 2011.

3.21 The Generative Topographic Mapping and Interactive Visualization

Ian Nabney (Aston University - Birmingham, GB)

Joint work of Nabney, Ian; Tino, Peter; Maniyar, Dharmesh; Schroeder, Martin

The Generative Topographic Mapping (GTM) is a probabilistic generative data model. Using Bayes' theorem, the mapping can be inverted and used for visualization. Because the model is a constrained mixture of Gaussians, an (extended) EM algorithm can be used to train models. The smooth mapping defined by GTM defines a two-dimensional manifold embedded in data space: geometric measures (e.g. magnification and curvature) can be visualized to understand the embedding and diagnose modelling flaws.

More recent advances include modelling missing data, discrete variables, and hierarchies: all of these can be handled in a consistent probabilistic framework. With a bit more analysis, it is possible to incorporate prior knowledge of variable correlation structure (with block-structured covariance models) and unsupervised feature selection (with minimum message length criteria). The talk concluded with a short demonstration of a visualization system that integrates machine learning and information visualisation (Data Visualization and Modelling System: DVMS) written in Matlab which is available from the Aston website. http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/

3.22 Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization

Jaakko Peltonen (Aalto University, FI)

Joint work of Venna, Jarkko; Peltonen, Jaakko; Nybo, Kristian; Aidos, Helena; Kaski, Samuel
Main reference Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization. Journal of Machine Learning Research, 11:451-490, 2010.
URL http://jmlr.csail.mit.edu/papers/v11/venna10a.html

Nonlinear dimensionality reduction methods are often used to visualize high-dimensional data, although the existing methods have been designed for other related tasks such as manifold learning. It has been difficult to assess the quality of visualizations since the task has not been well-defined. We give a rigorous definition for a specific visualization task, resulting in quantifiable goodness measures and new visualization methods. The task is information retrieval given the visualization: to find similar data based on the similarities shown on the display. The fundamental tradeoff between precision and recall of information retrieval can then be quantified in visualizations as well. The user needs to give the relative cost of missing similar points vs. retrieving dissimilar points, after which the total cost can be measured. We then introduce a new method NeRV (neighbor retrieval visualizer) which produces an optimal visualization by minimizing the cost. We further derive a variant for supervised visualization; class information is taken rigorously into account when computing the similarity relationships. We show empirically that the unsupervised version outperforms existing unsupervised dimensionality reduction methods in the visualization task, and the supervised version outperforms existing supervised methods.

3.23 Visualization of Learning Processes - A Problem Statement

Gabriele Peters (FernUniversität in Hagen, DE)

The visual representation of the results of machine learning algorithms can be regarded as an open research topic. But rather to restrict the visualization to only the results of machine learning approaches, the discussion should be expanded to the visualization of the learning processes themselves. Whereas a visualization of results promises a better interpretation of what has been learned, the visualization of learning processes may provide a better understanding of underlying principles of learning (also in biological systems).

Maybe it can also account for general insights in the possibilities of autonomous learning at all. In my talk I present briefly the architecture of a self-learning system with two levels of hierarchy together with some results obtained in a computer vision task. From this I derive questions of general interest such as possible options to visualize the flow of information in a dynamic learning system or the visualization of symbolic data.

3.24 Learning of short time series

Frank-Michael Schleif (Universität Bielefeld, DE)

Joint work of Schleif, Frank-Michael; Gisbrecht, Andrej; Hammer, Barbara
Main reference F.-M. Schleif, A. Gisbrecht, B. Hammer, Relevance learning for short high-dimensional time series in the life sciences. IJCNN'2012, 2012

The talk presented some concepts used to learn short and high dimensional time series.

Especially I detailed a method for topographic mapping and recent extensions thereof in the line of supervised relevance learning.

Challenges in the modeling and visualization were discussed.

3.25 Comparative Visual Cluster Analysis

Tobias Schreck (Universität Konstanz, DE)

Joint work of Schreck, Tobias; Tatu, Andrada; Maaß, Fabian; Bertini, Enrico; Keim, Daniel

Data that is to be analyzed with cluster analysis tools may be represented by sets of feature vectors stemming from alternative feature extraction processes.

Interesting cluster structures may reside in several of the alternative feature representations, and they may confirm, complement, or contradict each other. In this talk we consider the problem of comparative visual cluster analysis in multiple features spaces (or subspaces). We first briefly review a previously proposed method for visual comparison of multiple feature spaces represented by Self-Organizing Map models. We then discuss ongoing work that aims to make use of automatic subspace selection methods. First results based using the SURFING subspace selection method are reported. The basic idea is to define a custom similarity function for the subspaces. The function currently considers the intersection of the selected dimensions as well as the agreement in clustering structures exhibited in the subspaces. Different visual representations based on MDS layouts, TreeMap layouts etc. as well as interaction techniques are investigated. Eventually, our approach should help analysts in identifying the most interesting subspaces from a potentially much larger set of subspaces reported by the subspace selection method.

3.26 Visualization of (machine) learning processes and dynamic scenarios

Marc Strickert (Universität Marburg, DE)

The title can be related to an overwhelming plenitude of aspects such as functional brain imaging, motion sensor and eye tracker analysis, neural spike train observations, phase space portraits, or time series and data stream mining. To focus the wide topic on one essential commonality, this involves the transformation of spatio-temporal multi-dimensional input data into representations that are compatible with analysts' world view. This requires a compatibility between the data model and the world model mainly constituted by three spatial coordinates, color, intensity, and experience of spatio-temporal contiguity.

In machine learning methods sequential signals are often recursively mixed with a representation of the most recent internal state for modeling first-order context. The current model state is thus a representation of possibly unifying encodings of external dynamics. Depending on different readout functions applied to the model parameters different aspects of the input stream are focused on.

After all, structure detection, including ordering and convergence trends, is considered as crucial component for extracting aspects which are potentially relevant for visualization.

3.27 Prior knowledge for Visualization or prior visualization results for knowledge generation - a chicken-egg problem?

Holger Theisel (Universität Magdeburg, DE)

It seems that both communities - Machine Learning and Visualization - use different words for the same concept, and even use the same words for different concepts. This holds both for "prior knowledge" and "visualization". Being aware of this, the following questions are discussed: Are there new ML algorithms when the goal is preparation for a interactive visual analysis? Are there new Vis approaches when the goal is not complete insight but preparation of an automatic analysis?

3.28 Interactive decision trees and myriahedral maps

Jarke J. Van Wijk (TU Eindhoven, NL)

Joint work of Van Wijk, Jarke J.; van den Elzen, Stef;
Main reference Stef van den Elzen, Jarke J. van Wijk: BaobabView: Interactive construction and analysis of decision trees. IEEE VAST 2011: 151-160

In my talk, I present BaobabView, a system developed by Stef van den Elzen. It enables users to construct, inspect, and evaluate decision trees via a wide range of features. Next, myriahedral projections are presented via a video. These are mappings of the sphere to the plane using an approximation of the sphere with a large number of facets, which are cut and folded out. Finally, a short demo of SeifertView is given. Seifert surfaces are orientable surfaces that are bounded by knots or links. They illustrate that 2-manifolds embedded in 3D can take complex shapes.

3.29 Clustered graph, visualization, hierarchical visualization

Nathalie Villa-Vialaneix (Université Paris I, FR)

Joint work of Villa-Vialaneix, Nathalie; Rossi, Fabrice
Main reference Rossi & Villa-Vialaneix (2011) Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets. Journal de la Société Française de Statistique, 152, 34-65.
URL http://hal.archives-ouvertes.fr/hal-00651577/

Clustering is a useful approach to provide a simplified and meaningful representation of large graphs. By extracting dense communiites of nodes, the "big picture" of the network organizatin is enlighten. Moreover, hierarchical clustering may help the user to focus on some parts of the graph which is of interest for him and which can be displayed with finer and finer details.

This talk will try to present some open issues with graph visualization based on a hierarchical nodes clustering. These issues include displaying the clusters in a coherent way between the different layers of the hierarchy or integrating information about the clustering evaluation in the visualization. It is related to the article [1].

References

1 Rossi, F. and Villa-Vialaneix, N.
(2011) Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets. Journal de la Société Française de Statistique, 152, 34-65.

3.30 Perceptual Experiments for Visualization

Daniel Weiskopf (Universität Stuttgart, DE)

I briefly describe and discuss a few examples of user experiments that investigate the visual perception of visualization results, including studies that use methods from vision research, eye-tracking experiments in the context of the visualization of node-link diagrams of graphs and trees, as well as learning attention models for video visualization by utilizing eye- tracking data.

Background references:

References

1 Michael Burch, Corinna Vehlow, Natalia Konevtsova, Daniel Weiskopf: Evaluating Partially Drawn Links for Directed Graph Edges. *Graph Drawing 2011*, 226-237 (2011)

2 Michael Burch, Corinna Vehlow, Fabian Beck, Stephan Diehl, Daniel Weiskopf: Parallel Edge Splatting for Scalable Dynamic Graph Visualization. *IEEE Trans. Vis. Comput. Graph.* 17(12): 2344-2353 (2011)

3 Benjamin Höferlin, Hermann Pflüger, Markus Höferlin, Gunther Heidemann, Daniel Weiskopf: Learning a Visual Attention Model for Adaptive Fast-Forward in Video Surveillance. *Proceedings of International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 25-32 (2012).

3.31 Introduction to embedding

Laurens van der Maaten (TU Delft, NL)

In this talk, I presented an overview of some major embedding techniques and explained their strengths and weaknesses. In particular, I explained principal components analysis, locally linear embedding, and t-distributed stochastic neighbor embedding. In addition, I showed examples of embedding techniques that go beyond traditional dimensionality reduction and multidimensional scaling.

In particular, I covered embedding techniques thta learn representations from non-metric similarities such as word associations, co-occurrences, and partial order rankings.

4 Working Groups

4.1 Results of Working Group: Visualization of Dynamic Learning Processes

Biehl, Michael; Bunte, Kerstin; Peters, Gabriele; Strickert, Marc; Villmann, Thomas;

Main reference Thomas Leopold, Gabriele Kern-Isberner, and Gabriele Peters. Combining Reinforcement Learning and Belief Revision - A Learning System for Active Vision. 19th British Machine Vision Conference (BMVC 2008), edited by M.Everingham, Ch. Needham, and R. Fraile, Vol. 1, pp. 473-482, Leeds, UK, 2008.

We took the learning system [1] proposed by G. Peters in her talk "Visualization of Learning Processes - A Problem Statement" as example for a dynamic learning process and figured out which components can be visualized and by which means. The system has two learning levels: one with if-then rules (boolean expressions) and one with qualities of state- action pairs. Relevant questions to ask are: What are parameters of the system? Which parameters define states to be visualized? How to visualize the list of rules and the state- action table? How to visualize dynamic processes between the hierarchy levels? We proposed several means and solutions to answer these questions and intend to submit these considerations as position paper to a suitable conference or workshop.

References

1 T. Leopold and G. Kern-Isberner and G. Peters. *Combining Reinforcement Learning and Belief Revision - A Learning System for Active Vision.* BMVC, 2008

4.2 Results of Working Group: Model Visualization - Towards a Tight Integration of Machine Learning and Visualization

Mansmann, Florian; Schreck, Tobias; Come, Etienne; Wijk, Jarke van

Choosing and configuring an appropriate Machine Learning model to solve a given analysis task is crucial for arriving at useful results. Models in Machine Learning are potentially complex and sometimes hard to understand for non-experts, and often regarded and applied as black boxes. In this working group we discussed about approaches to 'opening' the black boxes by visualizing not only the data space, but also the space of model parameters. Our goal is to eventually arrive at better selection and configuration of Machine Learning models using interactive visualization. We started our discussion with the question 'What is a model?' and developed a draft reference model for Model Space Visualization. To this end, we built on existing process models, including the Information Visualization model of Card, MacKinley and Shneiderman and the Visual Analytics model proposed by Thomas and Keim. Our model adds one level of detail to the formalism and distinguishes between expert and user roles. In particular, this new process model makes the integration of Machine Learning and Visualization explicit. Consideration of model instances and parameter sets as part of the workflow in our model aims at a tighter integration of machine learning into the interactive analysis process. Also, the model is aiming as a reference structure to survey and classify existing works in Visual Analytics and Visual Data Mining. These latter points are seen as interesting future work.

4.3 Results of Working Group: Embedding techniques at the crossing of Machine Learning and Information Visualization

Aupetit, Michael; Lee, John;

4.3.1 Attendees

- Information Visualisation: D. Keim; L. Zhang
- Machine Learning: M. Verleysen; J.A.Lee; M. Aupetit; S. Kaski; J. Peltonen; L. van der Maaten; F.-M. Schleif

4.3.2 Emerging topics of interest

- from the Information Visualisation perspective
  Getting trust from the analyst is fondamental to make embedding common visual analytics tools. For this, the projection must not change drastically if no strong changes occur in the data distribution, so questions are:
  - How to make embeddings robust to noise and outliers?
  - How to make embeddings stable adding new data points and against local optima and different initialisation ?

  Interactivity is also a very important point in visual analytics,
  - How to deals with massive datasets in terms of speed and quantity of data to visualize?

  - How to link embeddings of different local subspaces to get better understanding of the data?

  Understandability is another main issue with non linear embeddings (axes have no sense):
  - How to connect embeddings to the meaning of the original data features?
- from the Machine Learning perspective

  Assessing Visual Analytic tools needs well defined tasks:
  - Can we define a taxonomy of tasks and data types that could benefit from embeddings?

  Raw data have to be preprocessed before embedding:
  - Which kind of preprocessing has to be done before embedding?
  - Which kind of similarity measures make embeddings more efficient for which kind of task?
  - How to deal with discrete, non-Cartesian, missing data?

4.3.3 Intended actions

- Sharing of various data sets (InfoVis) and embedding methods (ML)
- Build a joint InfoVis/ML taxonomy
- Organizing a workshop and a tutorial on embeddings at the next IEEE VisWeek 2012 conference from which we can edit a special journal issue on this topic

We thank all the contributors to this group, and all the colleagues from the Dagstuhl seminar for fruitful discussions.

4.4 Results of Working Group: Evaluation

Nabney, Ian; Cook, Di; Fisher, Gisbrecht, Andrej; Brian; Hagen, Hans, Hoffman,Heicke

URL http://db.tt/G0Ajn0V6

Our outcomes were captured by Ian Nabney in a mindmap which can be found at the URL below as a .mm file (http://db.tt/G0Ajn0V6). These files can be opened in a number of applications including Freemind http://freemind.sourceforge.net/

4.5 Results of Working Group: Fast Machine Learning

Joern Kohlhammer (Fraunhofer Inst. - Darmstadt, DE)

This session discussed the topic of "Fast ML for interactive visualization" and what the different perspectives are in ML and Infovis/VA/visualization.

It turned out that the ML community in its various sub-communities is not focused per se on performance issues of their algorithms, at least not to the extent of trying to achieve real-time capabilities. The InfoVis and VA community on the other hand is actively looking for high-performance, automated methods that can be coupled with visualization techniques to include more and more data in an interactive analysis. Response times are very important for interactive techniques and such response times do not play a major role in many ML approaches.

There were several thoughts about the user influence on ML methods, which might be beneficial to the ML community. One can distinguish between an internal coupling of methods, where the user interactively influences the automated methods during run-time, or an external coupling, which focuses on the flexible ensemble of ML methods and visualization methods along a structured analysis workflow.

The outcome of the discussion was that it would be highly interesting for the visualization community to learn about the current extent of research in this direction, i.e. a more performance-driven view on current research in ML:

- What type of ML methods do exist?
- Which sub-communities (or which research groups specifically) work on high-performance ML approaches?
- What are these approaches in detail? What are their characteristics, scalability constraints, data types, etc.?
- How could we jointly work on coupling such approaches with InfoVis and VA? Are their existing joint efforts, best practices, examples?
- What are the plans and future work in these Vis-relevant areas?

Next steps:

Our idea was to plan a tutorial for VisWeek 2012 with the tentative title "A performance perspective on machine learning for visualization", to be submitted by 30 April 2012. The tutorial could be a half day or full day tutorial, depending on the outcome of the next planning steps. There could be 4-5 speakers or even more, again depending on the structure.

The tutorial should give an overview of ML methods and go into detail on the high-performance methods (along the lines of the above questions), building a possible repertoire of ML methods for visualization.

The joint understanding is that the talks in the tutorial should be held by ML experts, but with strong involvement of visualization experts in the planning phase to make sure that the talks are targeted at and are adequate/educational for the visualization community.

4.6 Results of Working Group: Future analysis environments

Fekete, Jean-Daniel

The current situation of data analysis environments can be summarized simply by saying that there are numerous environments each of which has very satisfied users that do not want to switch to another solution. The only reasonable solution to avoid duplications of efforts is therefore to have some form of interoperability between environments. This can be provided at:

- a library level, with difficulties induced by differences between programming languages;
- an export/import level using e.g. xml formats with difficulties related to encoding and similar issues;
- a component level via rpc or web services mechanisms.

While interoperability might save the day, it has its share of problems:

- speed and latency;

- data duplication;
- limitation of some of the environments.

While one environment could rule them all, this seems unlikely, and improving interoperability seems a simpler goal. This needs not only the ability to share data, but also the support of notifications (of changes) and of metadata. One possible plan would be:

- specify and implement a sharing mechanism for R, Matlab, Excel...
  - how to connect/disconnect to a shared datatable
  - how to load content lazily
  - how to emit and receive notifications
  - how to manage content consistency
  - etc.
- test it

4.7 Results of Working Group: Structured/relational data

Nathalie Villa-Vialaneix

Structured and relational data have been discussed and several issues have been extracted:

- clustering issues: evaluating the quality/relevance of a clustering/cluster
- taking into account heteregeneous data: heteregeneous data could lead to different clusterings: put the user in the loop to help find a consensual user-driven clustering
- metric came from mathematics; use the users' suggestions to try to find a consensus among the human experts and use ML to extract a relevant metric that fits the users' suggestion
- how to find a relevant labelling for a cluster: give the user hints automatic help for labelling