Intelligent Data Analysis

13.08 - 18.08.2000

organized by

Michael R. Berthold, Rudolf Kruse, Xiaohui Liu, and Helena Szczerbicka

1 Introduction

For the last decade or so, the size of machine-readable data sets has increased dramatically and the problem of "data explosion" has become apparent. On the other hand, recent developments in computing have provided the basic infrastructure for fast access to vast amounts of online data and many of the advanced computational methods for extracting information from large quantities of data are beginning to mature. These developments have created a new range of problems and challenges for the analysts, as well as new opportunities for intelligent systems in data analysis and have led to the emergence of the field of Intelligent Data Analysis (IDA), a combination of diverse disciplines including Artificial Intelligence and Statistics in particular. These fields often complement each other: many statistical methods, particularly those for large data sets, rely on computation, but brute computing power is no substitute for statistical knowledge.
Although many interesting systems and applications have been developed in the field, much more needs to be done. For instance, most of the data collected so far have not been analyzed, and there are few tools around which allow the effective analysis of "'big data"'. Different analysis strategies may be applied to the same problem and it is often difficult to judge which is the most appropriate; and the division of the work between the computer and the analyst most effectively is by and large still work of art. Many different application areas offer vast potential for improvement, examples include such diverse fields as the automatic monitoring of patients in medicine (which requires an understanding of the underlying decision process), optimization of industrial processes, and also the extraction of expert knowledge from observations of their behavior.
To discuss these and many other related topics, the First International Symposium in Intelligent Data Analysis (IDA-95) was held in Baden-Baden, followed by the second IDA symposium (IDA-97) in London and the third (IDA-99) held in Amsterdam. All three symposia were successful in attracting a large number of people with different backgrounds including AI, pattern recognition and

statistics as well as people from diverse application fields. However, these large symposiums tend to lack of close interaction and discussion among people, which are relatively easy with a limited number of participants. Therefore the goal of this Dagstuhl seminar was to bring together a small number of experts from the various disciplines to discuss important issues in Intelligent Data Analysis, review current progress in the field, and identify those challenging and fruitful areas for further research.

2 Scope

The goal of this Dagstuhl Seminar was to focus on some key issues in intelligent data analysis that are directly relevant to the above aspects, both from the application and theoretical side:

- Strategies: Data analysis in a problem-solving context is typically an iterative process involving problem formulation, model building, and interpretation of the results. The question of how data analysis may be carried out effectively should lead us to having a close look not only at those individual components in the data analysis process, but also at the process as a whole, asking what would constitute a sensible analysis strategy. This strategy would describe the steps, decisions and actions which are taken during the process of analyzing data to build a model or answer a question. The work reported so far in this area has very much been restricted to relatively small problems. It is important to further our understanding of strategic issues in the context of large data-analysis tasks.

- Integration: In addition to careful thinking at every stage of an analysis process and intelligent application of relevant domain expertise regarding both data and subject matters, Intelligent Data Analysis requires critical assessment and selection of relevant analysis approaches. This often means a sensible integration of various techniques stemming from different disciplines, given that certain techniques from one field could improve a method from another one.

- Data Quality: Data are now viewed as a key organizational resource and the use of high-quality data for decision making has received increasing attention. It is commonly accepted that one of the most difficult and costly tasks in modern data analysis is trying to obtain clean and reliable data. There are so many things that can be wrong with real-world data: they can be noisy, incomplete and inconsistent, and it is not always easy to handle these problems. Since the use of "wrong" kind of data or very "low-quality" data often leads to useless analysis results, research on data quality has attracted

a significant amount of attention from different communities including information systems, management, computing, and statistics. Much progress has been made, but further work is urgently needed to come up with practical and effective methods for managing different kinds of data quality problems in large databases.

- Scalability: One of the key issues involved in large-scale data analysis is "scalability", e.g. if a method works well for a task involving a dozen variables, is it still going to perform well for one with over 100 or 1000 variables? The "'curse of dimensionality"' is still with us, and there has been a lot of intensive research on the development of efficient, approximate or heuristic algorithms that are able to scale well. For example, feature selection is one of the principal approaches to the scaling problem and a considerable amount of progress has been made in the field. However, many challenging issues still remain: how to develop a scalable algorithm when the data involved are highly-structured? and what would be the most effective way of sampling the most representative cases from a huge database for model-building? Currently, technical reports of analyzing "big data" are still sketchy. Analysis of big, opportunistic data (data collected for an unrelated purpose) is beset with many statistical pitfalls. We need to accumulate much more practical experience in analyzing large, complex real-world data sets in order to obtain a deep understanding of the IDA process.

3 Structure

Due to the interdisciplinary nature of the audience the real challenge of this meeting was the initiation of interactions across different disciplines. In order to provide the required background, one tutorial-style introduction was scheduled for each day, aiming to familiarize researchers with concepts from the various fields. These five surveys focused on some of the most important aspects of IDA:

- Generation, Representation, and Manipulation of Perceptions
- Association Rules
- Combining Classifiers
- Handling Missing Data
- Developing Mining Tools

The remainder of each day was used for presentations from the other attendees. On Friday we were fortunate to have Nicholas Deliyanakis from the Euopean Commisson join the seminar and present the new directions in funding under the 5^{th} Framework Programme.

4 Results

Results of this seminar are a better, deeper understanding of issues and methodologies in intelligent data analysis as well as insights into relevant real world problems. We hope this will lead to new interactions between disciplines, also initiated through insights into applications and the quite often very different emphasize they put on results and handling of the various techniques. Hopefully the proposed seminar has started new collaborations between people with different backgrounds, and has provided new insights regarding what is, or what is not, an intelligent way of analyzing data.

Contents

1 Introduction **1**

2 Scope **2**

3 Structure **3**

4 Results **4**

5 Survey: Generation, Representation, and Manipulation of Perceptions **6**

6 Survey: Mining and Using Association Rules **7**

7 Survey: Combination of Classifiers, A Liberal Overview **7**

8 Survey: Handling Missing Data in Data Mining **8**

9 Survey: Data Mining Systems Development **9**

10 Visualization of Fuzzy Models **10**

11 Fuzzy Clustering of Sampled Functions **11**

12 Genetic Fuzzy Systems for Intelligent Data Analysis **11**

13 Fuzzy Clustering in Intelligent Data Analysis **13**

14 Ensembles of Neural Networks **13**

15 Data Analysis with Neuro-Fuzzy Methods **14**

16 Statistical Edits - A Firewall against inconsistent Data Entry **15**

17 Intelligent Data Analysis of Patient and User Data **15**

18 Data-Driven Fuzzy Cluster Modelling **16**

19 Multi-relational datamining **18**

20 Hierarchical Classification **18**

21 Continuous Learning **19**

22 Clustering and Visualisation of Large and Highdimensional Data Sets **20**

23 Using Data Mining for Improving a Wafer Cleaning Process in the VLSI Industry **21**

24 The Funding of Science and Technology under the EC 5th Framework Programme **22**

5 Survey: Generation, Representation, and Manipulation of Perceptions

Enrique H. Ruspini[1]

The increased ability to access repositories of representations of complex objects, such as biological molecules or financial time series, has not been matched by the availability of tools that permit locating them, visualizing their characteristics, and describing them in terms that are close to the language and experience of the users of those data collections. The representation methods and organization schemes employed in the majority of these repositories are based, rather, on considerations related to computational simplicity and efficiency. The observation that humans typically resort to qualitative descriptions of complex objects to describe interesting features of these entities suggests that the automated generation of those descriptions might substantially improve the accessibility and usefulness of repositories of complex objects.

We present results of ongoing research on methods for the automatic derivation of qualitative descriptions of complex objects. The ultimate goals of these investigations are the development of a methodology for the qualitative representation of complex objects, the systematic search and retrieval of objects based on those representations, and the discovery of knowledge based on the study of collections of such qualitative descriptions.

We present a Pareto genetic algorithm for the identification of interesting object structures. This problem is posed as a multiobjective optimization problem involving solution quality and explanation-extent goals. The formulation of this problem is noteworthy because of its treatment of clustering as the isolation of individual clusters (i.e., as opposed to the determination of optimal partitions), its lack of assumptions about the number of clusters, and its ability to deal simultaneously with multiple types of interesting structures.

We will discuss also methods for the succinct description of structures—lying on the effective frontier—uncovered by our optimization algorithm. We rely also on clustering and summarization techniques to produce such a summarization. Finally, we will present experimental results of the application of our methodology to the qualitative description of financial time series.

[1]Joint work with Igor S. Zwir

6 Survey: Mining and Using Association Rules

Bing Liu

Association rule mining is an important model in data mining. Since it was first introduced in 1993, it has been studied extensively in the database and data mining community. Its mining algorithm discovers all item associations (or rules) in the data that satisfy the user-specified minimum support (minsup) and minimum confidence (minconf) constraints. Minsup controls the minimum number of data cases that a rule must cover. Minconf controls the predictive strength of the rule. In this talk, we focus on the following:

1. Introduce the classic association rule model and the most popular association rule mining algorithm, the Apriori algorithm.
2. Discuss a key shortcoming of the classic model, a single minimum support for the whole data set, which causes the rare item problem. We then introduce a more general model and algorithm for mining association rules with multiple minimum supports, which helps to solve the rare item problem.
3. Describe a key application of association rules, i.e., classification based on associations (or association rules based classification). Association rules can be used to build accuracy classifiers and scoring models.
4. Discuss post-processing of association rules. One problem with association rule mining is that it often produces a huge number of rules, which makes it very hard for the user to analyze to find those interesting ones. We discuss a rule summarization and organization technique to make it easier for the user.

7 Survey: Combination of Classifiers, A Liberal Overview

Lawrence O. Hall

It has been shown experimentally that the combination of classifiers can often provide higher accuracy results than can be obtained from any individual classifier. This talk describes how to choose classifiers for combination by examining

their bias and variance. It describes a subset of the combination techniques that have been shown to be effective, such as uniform weighted voting, weighted voting, bayesian combination, and learned combiners. Also, methods of generating multiple classifiers for combination are discussed. The choice of the train set to which each classifier is applied is discussed. How to generate multiple classifiers that can be effectively combined from a single underlying learning technique (which must be "unstable") is described. Examples of unstable classifiers, such as decision trees or rule learners or neural networks are enumerated. The techniques of bagging and boosting are discussed. Examples of the improvement obtained by combining classifiers are given.

8 Survey: Handling Missing Data in Data Mining

Ad Feelders

In many applications of data mining a - sometimes considerable - part of the data values is missing. Despite its frequent occurrence, most data mining algorithms handle missing data in a rather ad-hoc way, or simply ignore the problem. We discuss a number of ad-hoc methods for handling missing data (complete case analysis, single imputation) and illustrate their shortcomings. Then we describe two principled approaches, namely Expectation Maximization (EM) and model-based Multiple Imputation (MI).
First, we apply EM to the reject inference problem in credit scoring. Reject inference is the process of estimating the risk of defaulting for loan applicants that are rejected under the bank's current acceptance policy. This can be viewed as a missing data problem since the outcome (defaulted or not) of the rejected applicants is not observed. We propose a new reject inference method based on mixture modeling, that allows the meaningful inclusion of the rejects in the estimation process. We describe how such a model can be estimated using the EM-algorithm. An experimental study shows that inclusion of the rejects can lead to a substantial improvement of the resulting classification rule.
Second, we investigate model-based Multiple Imputation, which is based on filling-in (imputing) one or more plausible values for the missing data. One advantage of this approach is that the imputation phase is separated from the analysis phase, allowing for different data mining algorithms to be applied to the completed data sets. We compare the use of imputation to surrogate splits, such as used in CART, to handle missing data in tree-based mining algorithms.

Experiments show that imputation tends to outperform surrogate splits in terms of predictive accuracy of the resulting models. Averaging over $M > 1$ models resulting from M imputations yields even better results as it profits from variance reduction in much the same way as procedures such as bagging.

9 Survey: Data Mining Systems Development

Arno Siebes

In this talk I gave an overview of my experiences in developing data mining systems, both a research prototype and a commercial systems. Luckily, the development of these systems started in together, their ways diverted, however, because the needs in a commercial environment differ substantially from those in a research environment.
I focussed the discussion of design decisions on two aspects, viz., the mining engine and database aspects.
Data mining algorithms consist of just a few components. Firstly, the representation language, the set of all possible models. Secondly, a quality function that determines how well a model fits (part of) the database. Thirdly a search algorithm, which consists of a general strategy (such as hill-climbing or genetic search) and operators that implement the strategy on a particular representation language (such as a neighbour operator for hill-climbing).
The architecture faithfully reflects this decomposition of algorithms with a search manager (for the search strategies), a model generator (for the search operators) and a quality computer (for the quality function); only the last component is allowed to query the database. The communication between these components is via a special database: the search space manager, which stores the search space as far as it is explored. This latter aspect allows a user to track the search and interact (steer) with it.
The components-based approach did not allow for as much code-reuse as we hoped, but it did make the development more manageable. Moreover, one point for database access did prove a success. All the more, by restricting the database queries by the quality computer to data-cubes.
The KESO system has its own internal DBMS, viz., Monet a joint effort of CWI and the University of Amsterdam. Monet is a main memory database system that relies on decomposed data storage (i.e., per attribute rather than per row). Given that many mining algorithms zoom in on the data, this allows for optimization. More optimization opportunities were given by the restriction

to data-cube computation and the fact that mining algorithms generate batches of related queries at the time. Multi-query optimization of such batches gave again a considerable speed-up.
Data Surveyor from Data Distilleries shares these architectural with the KESO system. Still, developing a commercial mining system posed problems that the KESO development did not have. The two most important ones are in the interface and in the goal of the system.
In a commercial environment, ease of use is of crucial importance. For example, one should not take it for granted that a user is good at SQL programming. Rather, to extract the relevant information from various data sources, the user should be given an intuitive interface. Similar observations hold for aspects such as the derivation of new attributes, the parameter settings of the mining algorithms etc.
An even more important observation is that it is very difficult to sell a data mining system as such. It is far easier to market a "vertical solution" for an area such as Customer Relationship Management. The fact that there is a (the clients do not need or want more) data mining algorithm hidden in this software should (almost) be hidden from the user.

10 Visualization of Fuzzy Models

Michael R. Berthold

In this talk we present two approaches to visualize a potentially high-dimensional and large number of (fuzzy) rules in two dimensions.
This first visualization presents the entire set of rules to the user as one coherent picture. We use a gradient descent based algorithm to generate a 2D-view of the rule set which minimizes the error on the pair-wise fuzzy distances between all rules. This approach is superior to a simple projection and also most non-linear transformations in that it concentrates on the important feature, that is the inter-point distances. In order to make use of the uncertain nature of the underlying fuzzy rules, a new fuzzy distance-measure was developed.
The second approach is based on parallel coordinates which allow to visualize data in two dimensions. Essentially, parallel coordinates transforms multi-dimensional patterns into two-dimensional patterns without loss of information. Visualization is facilitated by viewing the two-dimensional representation of the n-dimensional data points as lines crossing n parallel axes, each of which represents one dimension of the original feature space. This approach scales well with increasing n

and has already been incorporated in some data analysis tools. We additionally present how fuzzy points can be visulized using this technique.
The visualizations of a rule set for the well-known IRIS dataset as well as fuzzy models for other benchmark data sets are illustrated and discussed.

11 Fuzzy Clustering of Sampled Functions

Frank Höppner[2]

Fuzzy clustering algorithms perform cluster analysis on a data set that consists of feature attribute vectors. In the context of multiple sampled functions, a set of samples (e.g. a time series) becomes a single datum. We show how the already known algorithms can be used to perform fuzzy cluster analysis on this kind of data sets by replacing the conventional prototypes with sets of prototypes. This approach allows reusing the known algorithms and works also with other data than sampled functions. Furthermore, to reduce the computational costs in case of single-input/single-output functions we present a new fuzzy clustering algorithm, which uses for the first time a more complex input data type (data points aggregated to data-lines instead of raw data). The new alternating optimisation algorithm performs cluster analysis directly on this more compact representation of the sampled functions.

12 Genetic Fuzzy Systems for Intelligent Data Analysis

Frank Hoffmann

In my talk I present a new approach to structure identification of Takagi-Sugeno-Kang (TSK) Fuzzy Models. A fuzzy rule constitutes a local linear model, that is valid for the region described by the premise part of the rule. We employ Genetic Programming (GP) to find an optimal partition of the input space into

[2]This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant no. Kl 648/1-1.

Gaussian, axis-orthogonal fuzzy sets. For a given partition of clusters, a local weighted least squares algorithm estimates the linear parameters in the rule conclusion. The non-homogeneous distribution of clusters enables a finer granulation in regions where the output depends in a highly non-linear fashion on the input. We compare the GP approach with a greedy partition algorithm (LOLIMOT) for modeling an engine characteristic map.

Fuzzy systems are particular intriguing for intelligent data analysis, as there mode of approximate reasoning resembles the decision making employed by humans. A fuzzy system is often designed by interviewing an expert and formulating her implicit knowledge of the underlying process into a set of linguistic variables and fuzzy rules. An alternative is to automatically extract the fuzzy model from training data by means of a learning process.

Evolutionary algorithms are optimization methods that imitate the processes that occur in natural evolution and genetics. They maintain a population of candidate solutions which evolves over time by means of genetic operators such as mutation, recombination and selection. Genetic Programming (GP) is concerned with the automatic generation of computer programs by evaluating their performance against a set of training cases.

Combining evolutionary algorithms with fuzzy modeling, results in so called genetic fuzzy systems, that either tune or learn a set of fuzzy rules and membership functions from data. In our case the tree structure processed by the GP generates a partition of the input space into clusters. Each cluster corresponds to a fuzzy rule, which describes the relationship between input and output for the region around the cluster center. The GP performs the structure identification of the model, namely the distribution of cluster centers, whereas the weighted local least squares handles the parameter estimation part as it computes the optimal linear parameters in the rule conclusion.

We compared the GP approach with a greedy heuristic, called local linear model trees (LOLIMOT) for the task of modeling an engine characteristic map. For fuzzy systems with a few number of models the GP was able to find partitions with slightly improved model accuracy compared to the greedy LOLIMOT algorithm. On the other hand, GP tends to be less robust than LOLIMOT in the sense that it does not consistently finds good partitions, especially as the number of rules increases. GP requires a substantially larger amount of computation time compared to the sequential, very efficient LOLIMOT algorithm. This computational burden can only be justified in applications for which small improvements of accuracy of the model carry a large benefit.

13 Fuzzy Clustering in Intelligent Data Analysis

Frank Klawonn

Fuzzy control at the executive level can be interpreted as an approximation technique for a control function based on typical, imprecisely specified input-output tuples that are represented by fuzzy sets. The imprecision is characterized by similarity relations that are induced by transformations of the canonical distance function between real numbers. Taking this interpretation of fuzzy controllers into account, in order to derive a fuzzy controller from observed data typical input-output tuples have to be identified. In addition, a concept of similarity based on a transformations of the canonical distance is needed in order to characterize the typical input-output tuples by suitable fuzzy sets.
A variety of fuzzy clustering algorithms exists that are exactly working in this spirit: They identify prototypes and assign fuzzy sets to the prototypes on the basis of a suitable transformed distance. In this paper we discuss how such fuzzy clustering techniques can be applied to construct a fuzzy controller from data and introduce special clustering algorithms that are tailored for this problem.

14 Ensembles of Neural Networks

Joost N. Kok

We presented three ideas [1,2,3] concerning ensembles of neural networks.
The first idea is about the mathematical foundation of choosing the weights. If one chooses them according to the principle of maximum likelihood (using the assumption that the errors of the networks are normally distributed), then one should optimize $w\Sigma w^T$, where w is the vector of weights and Σ the covariance matrix under the constraint that the weights w sum to one. The interesting thing is (besides the nice mathematical foundation) that one gets rid of the constraint that all the weight should be positive. A closed form was given, but this uses the inverse of the covariance matrix and also some approximations were discussed.
The second idea is to train the neural networks such that they minimize the mean squared error but also such that they are not too much equal. In order to do this, a new energy funtion was given and a learning rule was derived.

The third idea was to apply Bayes theorem to predict new members of time sequences. If we take as prior the naive prediction (i.e. a normal distribution around the last value of the time sequence) and if the networks have normally distributed errors, then one can compute a posterior distribution which is also normally distributed.

References

[1] M.C. van Wezel, J.N. Kok and W.A. Kosters, Maximum likelihood weights for a linear ensemble of regression neural networks. In Proceedings ICONIP 98, Japan, 1998, pp. 498-501.

[2] M.C. van Wezel, M.D. Out and W.A. Kosters, Ensembles of Nonconformist Neural Networks, BNAIC'00, 2000.

[3] M.D. Out and W.A. Kosters, A Bayesian Approach to Combined Neural Networks Forecasting, presented at ESANN'2000 (The 8th European Symposium on Artificial Neural Networks), Brugge, April 26/28, 2000 (Proceedings pp. 323-328).

15 Data Analysis with Neuro-Fuzzy Methods

Rudolf Kruse

Fuzzy methods can be used to model linguistic terms, i.e. expressions of human language like large, small, hot, etc. Most fuzzy systems are based on if-then-rules (e.g. "if x is small then y is approximately zero"). Nowadays fuzzy systems are - in addition to control applications - applied in data analysis problems like classification, function approximation or time series prediction. Thier advantage is that they can provide simple intuitive models for interpretation and prediction. A typical industial application is the automatic gear box of the Volkswagen New Beetle, that uses fuzzy methods for the classification of drivers behaviour. In this talk the problem of learning fuzzy rules automatically from data is addressed. One solution is to use learning techniques derived from neural network theory in order to create a so called neuro-fuzzy system. The learnings algorithms in such a combined system are especially designed for their capability to produce interpretable fuzzy systems. It is important that during learning the main advantages of a fuzzy system- its simplicity and interpretability - do not get lost. A systematic overview of neuro-fuzzy methods for exploratory data analysis is given. A neuro-fuzzy architecture with four layers is analyzed in detail. This systems turned out to be useful in the context of the prediction of the German

DAX stock index. Finally the new Java-version of the NEFCLASS system is presented as an example for constructing a neuro-fuzzy system for data analysis in the context of classification.

16 Statistical Edits - A Firewall against inconsistent Data Entry

Hans-J. Lenz

Inconsistent data which enters into a database can corrupt a complete database, especially if OLAP (on-line analytical processing) is in use. Moreover, it is more efficient to check data at the entry instead of continuous checking a complete database, which is a NP-complete problem. Such a firewall consists of edits E, which represent semantically based rules, and can be viewed as mappings $E : S \leftarrow D$ from a multi-dimensional data or sampling space S to a decison space $D = \{0, 1, 2\}$ or $D = B(R)$. The decisions of a 'editor' or of the firewall are 'acceptance' ($d = 0$), 'rejection' ($d = 1$) or 'no decision' ($d = 2$). Alternatively, they may be intervals from the Borel set B defined on the set of real numbers R. The edits obey a flat onthology (taxonomy) and incorporate simple, logical, probabilistic, statistical (under a Gaussian regime) and fuzzy edits. The various edits are formally described, theorems and adjunct algorithms presented and the firewall functionality is illustrated by real life examples.

17 Intelligent Data Analysis of Patient and User Data

Rainer Malaka

Medical data often consist of very different data types such as patient age, sex, medical treatment and physiological measurements. Typically, data sets can have missing values. This makes it difficult to apply standard techniques of data analysis. Similarly, data from users of computer systems consist of heterogeneous data types. In a study, two data sets, one from the medical domain and one from

the usability domain, have been investigated using neural networks and various methods of preprocessing. The medical data set reflects patient data that undergo surgery showing if they suffer from post-operative nausea and vomiting (PONV). The user data set was derived from psychological experiments investigating the color preferences of user interfaces of software programs. Both data sets share some properties. They consist of 1000 up to 2000 patterns and each pattern has up to 100 features. For a classification of patients or users the high number of input variables may lead to over-fitting effects because of too many parameters that occur already in the input layer of a classifying neural network. In a comparison of three methods for input variable selection (logistic regression, information entropy and Spearmann's correlation coefficient), it could be shown that logistic regression is a powerful tool for reducing the input dimension and thus allowing for robust data analysis.

18 Data-Driven Fuzzy Cluster Modelling

Susana Nascimento

In partitional fuzzy clustering, like the fuzzy c-means method [1,2] and most of its extensions, membership functions are defined based on a distance function, and membership degrees express proximities of entities to cluster centers (i.e., prototypes). By choosing a suitable distance function different cluster shapes can be identified. However, with these methods, it may be difficult sometimes to explicitly describe how the fuzzy cluster structure found relates to the data from which it had been generated.
To provide a feedback from a cluster structure to the data, we employ a framework based on the assumption that the data are generated according to a cluster structure. The structure underlies the data in the format of the traditional statistical equation:

$$observed\ data = model\ data + noise.$$

In statistics, such an equation is accompanied by a probabilistic model of the noise. In our case, however, the model is not prespecified but rather derived from the data. This is why we concentrate on the model data part and leave the 'noise' to be considered as just the set of differences between the observed and model data. The differences thus are residuals rather than noise; they should be made as small as possible by fitting the model. This approach belongs to the more general framework of the so-called *approximation structuring* [3].

In the clustering model, we assume the existence of some prototypes, offered by the knowledge domain, that serve as "ideal" patterns to data entities. To relate the prototypes to observations, we assume that the observed entities share parts of the prototypes. It is these parts that constitute the model data. The underlying structure of this model can be described by a fuzzy c-partition defined in such a way that the membership of an entity to a cluster expresses *proportion* of the cluster's prototype reflected in the entity (*proportional membership*). This way, the underlying structure is substantiated in the fitting of data to the "ideal" patterns. This approach will be referred as *data-driven fuzzy cluster modelling*.
Distinct forms of pertaining observed entities to the prototypes lead to different instantiations of the data-driven fuzzy cluster modelling. In the ideal type fuzzy clustering model, for instance, any entity is represented as a convex combination of the prototypes; the convex combination coefficients are considered as the entity membership values. However, this model invokes the extremal rather than averaged properties of the data, which may lead to unrealistic solutions and makes the ideal type model much different from the other fuzzy clustering techniques.
In order to get solutions found with the model-based approach closer to those found with traditional fuzzy clustering techniques, we consider here a different way for associating observed entities and the prototypes. It is assumed that any entity may independently relate to any prototype, which is similar to the assumption underlying the fuzzy c-means criterion. The model is referred to as the *fuzzy clustering with proportional membership*, FCPM for short [4].
The results of an experimental study of various versions of FCPM and the fuzzy c-means algorithm (FCM) are presented and discussed, pointing out how the FCPM method fits the underlying clustering model, as well as its comparison with the fuzzy c-means.
As advantages of the data-driven cluster modelling framework, the following aspects can be pointed out. The assumption of a clustering model underlying the generation of data, implies direct interpretability of the fuzzy membership values. The ability to reconstruct the data from a clustering model is also a characteristic of this approach.
In particular, the FCPM approach seems appealing in the sense that, on doing cluster analysis, the experts of a knowledge domain usually have a conceptual understanding of how the domain is organized in terms of prototypes. This knowledge, put into the format of tentative prototypes, may well serve as the initial setting for data based structurization of the domain. In such a case, the belongingness of data entities to clusters are based on how much they share the features of corresponding prototypes. This seems useful in such application areas as mental disorders in psychiatry or consumer behavior in marketing.

References

[1] J. Dunn, A Fuzzy Relative of the Isodata Process and its Use in Detecting Compact, Well-Separated Clusters, Journal of Cybernetics, 3(3), pp. 32–57, 1973.
[2] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[3] B. Mirkin, Least-Squares Structuring, Clustering and Data Processing Issues. The Computer Journal, **41**(8), pp. 518-536, 1998.
[4] S. Nascimento, B. Mirkin and F. Moura-Pires, A Fuzzy Clustering Model of Data with Proportional Membership. The 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2000), IEEE , pp. 261-266, 2000.

19 Multi-relational datamining

Siegfried Nijssen

Most data mining algorithms have as input one single table, as for example the attribute-value learning algorithms. However, it is not always possible to merge a database of multiple tables into one table without losing or duplicating information. Special algorithms are needed for these databases. In my talk I discussed such special algorithms for finding association rules. They are based upon the so-called frequent query, which is a first-order logic extension of the frequent item set. To find these queries, previous frequent query discovery engines relied on ILP, which had a major impact on the performance of these algorithms. I argued that predicates should only be considered as views on a relational database and I described a tree datastructure which allows the discovery process to be performed reasonably fast in that case.

20 Hierarchical Classification

Guenther Palm

The talk focusses on two types of hierarchical neural networks for classification:

devision hierarchies and fusion hierarchies. Binary devision hierarchies can be built by 2-means clustering of the data, then sorting the class-labels accordingly, then training a network on this dichotomy. This process is repeated in each subclass until the complete classification is obtained. We considered RBF-SVM (radial basis function support vector machine) learning in this case and showed that the performance can be better (on some data) than other multiclass SVM procedures.
k-ary devision hierarchies were generated by first training a network on all data, then determining 'confusion-classes' of labels from the confusion matrix (i.e. all labels that were confused with a probability greater than d). The data from each of these confusion classes are then used to train a new network and so on. In each stage the confusion errors should become smaller. The procedure is stopped when no further improvement is obtained.
Fusion hierarchies naturally arise in sensor data fusion when the data are combined from many different sources or sensors. We considered only a very simple example with two sensors (lip-reading). Again a hierarchical approach may be better than training just one network on all data taken together. The reason for this becomes apparent when the different sensors are subject to varying and different amounts of noise. In this case it will often be more appropriate to combine the decisions of networks trained on the individual sensors than to combine the data from the sensors early on in one network. This was examplified on a data-set containing video-images of the mouth region and the acoustic signals generated when single letters were pronounced.

21 Continuous Learning

Peter Protzel

Typically, learning (e.g. in neural networks) takes place only during a certain time interval (training period), after which the system is put into operation with no further learning. In contrast, by "continuous learning" we mean that learning takes place all the time and is not interrupted. There is no difference between periods of training and operation, learning AND operation start with the first pattern.
The research on continuous learning with neural networks is motivated by industrial applications in process control with the task of approximating nonlinear, timevariant functions. The talk discusses the problems and implications of continuous learning (stability vs. plasticity dilemma) and introduces a new in-

cremental algorithm for the construction of a hybrid neural network architecture that is especially suited for continuous learning tasks.

22 Clustering and Visualisation of Large and Highdimensional Data Sets

Friedhelm Schwenker

In many practical applications one has to explore the underlying structure of a large set objects. Typically, each object is represented by a feature vector $x \in \mathbf{R}^d$. This data analysis problem can be tackled utilizing *clustering methods*. Here the aim is to reduce a set of M data points $X = \{x_1, \ldots, x_M\} \subset \mathbf{R}^d$ into a few, but representative prototypes or cluster centers $\{c_1, \ldots, c_k\} \subset \mathbf{R}^d$.
A neural network algorithm for clustering is Kohonen's *selforganizing feature map (SOM)* which is similar to the classical sequential k-means algorithm with the difference that in SOM the cluster centers are mapped to a fixed location of a 2- or 3-dimensional grid. The idea of this grid is that cluster centers corresponding to nearby points in the grid have nearby locations in the feature space, so Kohonen's SOM is able to combine clustering and visualisation aspects.
Another approach for getting an overview over a highdimensional data set $X \subset \mathbf{R}^d$ are *visualisation methods*. In multivariate statistics several linear and nonlinear techniques have been developed. A widely used nonlinear visualisation method is *multidimensional scaling (MDS)*. MDS is a class of distance preserving mappings from the data set X into a low-dimensional *projection space* $\mathbf{R}^r$. Each feature vector $x_i \in X$ is mapped to a point $y_i \in \mathbf{R}^r$ in such a way that the distance matrix $D_{\mathbf{R}^d} := (d_{\mathbf{R}^d}(x_i, x_j))_{1 \le i,j \le M}$ in feature space $\mathbf{R}^d$ is approximated by the distance matrix $D_{\mathbf{R}^r} := (d_{\mathbf{R}^r}(y_i, y_j))_{1 \le i,j \le M}$ in the projection space $\mathbf{R}^r$.
We describe an algorithm for exploratory data analysis which combines clustering and MDS. It can be used for the online visualisation of clustering processes in order to get an overview over large and highdimensional data sets. The algorithm, which is derived from multivariate statistical algorithms, may be considered as an alternative approach to the heuristic SOM. The adaptivity of this procedure makes it useful for many applications, where the clustering itself is part of a larger program operating in an enviroment under human supervision. Furthermore, it can help in forming hypotheses about the data, which in turn may be substantiated by other statistical methods.

23 Using Data Mining for Improving a Wafer Cleaning Process in the VLSI Industry

Armin Shmilovici

As the critical dimension in the Very Large Scale Industry (VLSI) decreased below the one micron level, the contamination of wafers with submicron particles became a major factor in the reduction of the yield. Typically, there are 6-9 particles per wafer, which can not be avoided. The current solution to that problem - wet cleaning takes a relatively lon time and involves thr use of highly toxic materials. Oramir is a machine building company located in Israel which developed a dry cleaning process which involves the use of a focused laser beam. The dry cleaning operates as follows: A measurement system is used to locate the positions of the contaminating particles and a laser beam is directed to each particle and fires a series of shots on each particle. A mixture of gases above the wafer is activated by the laser beam, and reacts with the particle so that it will detach his chemical links with the wafer and flow outside the wafer area. This process is extremely fast. Unfortunately, the reaction between the laser, the gases and the particle involves complex nonlinear processes, such as thermal shock, and presure waves. Modeling the reaction with first principle equations failed to provide a satisfactory model that will be used to calibrate the properties of the cleaning process. Oramir provided us with the results of about 200 experiments which were performed in different working conditions, and with the results of a linear multidimensional regression model. The linear model was not able to give accurate predictions of the process properties. There were 11 input variables (features) related to the properties of the process (e.g. power of the laser, gas concentrations); and two output variables %MOVAL - the percent of particles that were moved in an experiment from its original position, and %REMOVAL - the percent of particles that were removed succesfully from the wafer area. In order for the dry process to be economically effective for the VLSI manufacturers, we were asked to predict the input parameters of the cleaning process such that the %Removal is above 85%, and no damage is caused to the wafer during the cleanning process. It was decided to use 4 different data mining methods to test the problem, since the small number of observation did not guarantee that any of the mothods could produce favorable results. Experiments with a multipercepteron Neural Network were conducted with the gradient descent algorithm. The data were partitioned into two classes, the above 85% (GOOD) and the below 85% (BAD). After many experiments with different numbers of neurons, it was evident that the error rates in predicting the output variables, or in the classification of the experiment setting was unacceptable low - about 67%. Experiments with the C5.0 decision trees provided better results: a classification

error of 6.6% and a small number of rules were identified, that indicated that the variables related to the power of the laser were most significant, and also one of the gases was recommended for removal, since it was responsible for damaging the wafers during the cleaning process. The validity of the rules as recognized by the company expert. Experiments with reconstructability analysis of the data produced also good results - the input variables were ranked according to their desired influence in producing the best %REMOVAL, and the prediction of the output variables was above 88%. Experiments were made with composite classifiers. That is a second level neural net was used to generalize and enhance the predictions made from a mixture of 3 different first level classifiers. The stack generalization algorithm was used. A composite classifier of the above 3 different algorithms was able to reduce the classification error by up to 4%, to above 88%. We concluded from this experiment that the C5.0 decision tree classifier provided the most understandble and easy to use results, so its rules were recommended for implementation.

24 The Funding of Science and Technology under the EC 5th Framework Programme

Nicholas Deliyanakis

The Fifth Framework Programme (FP5), adopted on 22nd December 1998, defines the Community activities in the field of research, technological development and demonstration (RTD) for the period 1998-2002. It differs from its predecessors in that it has been conceived to solve problems and to respond to major socio-economic challenges facing the European Union. It focuses on a limited number of objectives and areas combining technological, industrial, economic, social and cultural aspects. It consists of seven Specific Programmes, four Thematic Programmes and three Horizontal Programmes:
The Thematic Programmes are:

- Quality of life and management of living resources
- User-friendly information society
- Competitive and sustainable growth
- Energy, environment and sustainable development.

The Horizontal Programmes, so called because they cover a wide range of scientific and technological fields, are:

- Confirming the international role of Community research
- Promotion of innovation and encouragement of participation of small and medium-sized enterprises (SMEs)
- Improving human research potential and the socio-economic knowledge base.

The programme is implemented through shared-cost activities; training (Marie Curie) fellowships; training and thematic networks; concerted actions; and accompanying measures.
The current seminar is one of the many scientific events (EuroConferences, Euro Summer Schools and others) supported by the High-Level Scientific Conferences activity, a part of the Human Potential Programme above. The general aims of this programme are to provide young researchers with opportunities for training; to sponsor international access to large-scale facilities; to promote scientific and technological excellence; to support policy making on research; and enhance the understanding of socio-economic issues. This particular activity, with a strong emphasis on training in all research areas, funds about 350 events every year, benefiting directly or indirectly of the order of 100,000 researchers over the duration of FP5.
The European Research Area is an initiative introduced by the European Commission, whose aim is to promote a real European research policy, through, for instance, networking, the creation of "virtual" laboratories, co-ordination of national and European research, and the promotion of mobility of researchers. Many possible solutions are being studied, and these will be implemented partly through the next Framework Programme, but also in several other ways.