Preface

Introduction

The dramatic increase in available computer storage capacity over the last 10 years has led to the creation of very large databases of scientific and commercial information. The need to analyze these masses of data has led to the evolution of the new field knowledge discovery in databases (KDD) at the intersection of machine learning, statistics and database technology. Being interdisciplinary by nature, the field offers the opportunity to combine the expertise of different fields into a common objective. Moreover, within each field diverse methods have been developed and justified with respect to different quality criteria. We have to investigate how these methods can contribute to solving the problem of KDD.

Traditionally, KDD was seeking to find global models for the data that explain most of the instances of the database and describe the general structure of the data. Examples are statistical time series models, cluster models, logic programs with high coverage or classification models like decision trees or linear decision functions. In practice, though, the use of these models often is very limited, because global models tend to find only the obvious patterns in the data, which domain experts already are aware of[1]. What is really of interest to the users are the local patterns that deviate from the already-known background knowledge. David Hand, who organized a workshop in 2002, proposed the new field of local patterns.

The Dagstuhl Seminar in April 2004 on Local Pattern Detection brought together experts from Europe, Japan, and the United States – 13 countries were represented. Moreover, the participants brought with them expertise in the following fields: decision trees, regression methods, bayesian models, kernel methods, inductive logic programming, deductive databases, constraint propagation, time series analysis, query optimization, outlier detection, frequent set mining, and subgroup detection. All talks were focused on the topic of local patterns in order to come to a clearer view of this new field.

Novelty of Local Pattern Detection

Researchers have investigated global models for a long time in statistics and machine learning. The database community has inspected the storage and retrieval of very large datasets. When statistical methods encounter the extremely large amount of records and the high dimensionality of the stored observations, exploratory methods failed. Machine learning already scales up to build up global

[1] I. Guyon, N. Matic and V. Vapnik. Discovering informative patterns and data cleaning. In *Advances in Knowledge Discovery and Data Mining* (pp. 181–204). AAAI Press/MIT Press, 1996.

models, either in the form of complete decision functions or in the form of learning all valid rules from databases. However, the classification does not deliver new, surprising insights into the data, and the valid rules reflect almost exactly the domain knowledge of the database designers. In contrast, what users expect from the exploratory analysis of databases are new insights into their data. Hence, the matter of interestingness has become a key issue. The success of Apriori or subsequently frequent set mining can be explained by it being the first step into the direction of local patterns. The correlation of more than the few features, which standard statistics could analyze, could successfully be determined by frequent set mining. Frequent set mining already outputs local patterns. Current research tasks within this set of methods include algorithmic concerns as well as the issues of interestingness measures and redundancy prevention. The collaboration of database specialists and data miners has led to the notion of inductive databases. The new approach writes measures of interest and the prevention of redundancy in terms of constraints. Also users can formulate their interests in terms of constraints. The constraints are pushed into the search process. This new approach was discussed at the seminar intensively and a view was found that covered diverse aspects of local patterns, namely their internal structure and the subjective part of interestingness as given by users.

Not all the exciting talks and contributions made their way into this book, particularly when a version of the talk was published elsewhere:

- Rosa Meo presented a language for inductive queries expressing constraints in the framework of frequent set mining.
- Bart Goethals offered a new constraint on the patterns, namely that of the database containing the minimal number of tiles, where each tile has the maximal number of '1'.
- Stefan Wrobel gave an in-depth talk on subgroup discovery, where he clearly indicated the problem of false discoveries and presented two approaches: the MIDOS algorithm, which finds subgroups according to the true deviation, and a sequential sampling algorithm, GSS, which makes subgroup discovery fast. He also tackled the redundancy problem by maximum entropy suppression effectively. Applications on spatial subgroup discovery concluded the talk.
- Arno Siebes employed a graphical view on data and patterns to express this internal structure. Moreover, aggregate functions along paths in these graphs were used to compute new features.
- Helena Ahonen-Myka gave an overview of sequence discovery with a focus on applications on text.
- Xiaohui Liu explained how to build a noise model using supervised machine learning methods and detect local patterns on this basis. Testing them against the noise model yields clean data. The approach was illustrated with two biomedical applications.
- Thorsten Joachims investigated internal structures such as parse-trees and co-reference pairing. He presented a general method for how such structures can be analyzed by SVMs. Moreover he showed how the combinatorial ex-

plosion of the number of constraints can be controlled by the upper bounds derived from statistical learning theory.

The book then covers frequent set mining in the following chapters:

- Francesco Bonchi and Fosca Giannotti show the use of constraints within the search for local patterns.
- Jean-Francois Boulicaut applies frequent set mining to gene expression data by exploiting Galois operators and mining bi-sets, which link situations and genes.
- Cline Rouveirol reports on the combination of frequent sets found in gene expression and genome alteration data.

Subgroup discovery is represented by three chapters:

- Nada Lavrac reports on successful applications of subgroup mining in medicine.
- Josef Fürnkranz presents a unifying view of diverse evaluation measures.
- Einoshin Suzuki investigates evaluation measures in order to distinguish local patterns from noise.
- Martin Scholz identifies global models with prior knowledge and local patterns with further, unexpected regularities. His subgroup discovery exploits iteratively a knowledge-based sampling method.

The statistical view is presented in the following chapters:

- Niall Adams and David Hand distinguish two stages in pattern discovery
  1. identify potential patterns (given a suitable definition);
  2. among these, identify significant (in some sense) patterns (expert or automatic).

  They notice that the former is primarily algorithmic and the latter has the potential to be statistical. They illustrate this with an application on discovering cheating students.
- Frank Höppner discusses the similarities and differences between clustering and pattern discovery. In particular he shows how interesting patterns can be found by the clever use of a hierarchical clustering algorithm.
- Stefan Rüping introduces a general framework in which local patterns being produced by different processes are identified using a hidden variable. This allows for the use of the EM algorithm to discover the local patterns directly, that is, without reference to the global data distribution. A new scaling algorithm handles the combination of classifiers. The method is illustrated using business cycle data.

Phenomena of time have always been of interest in KDD, ranging from time series analysis to episode learning. Here, three chapters are devoted to time phenomena:

- Claus Weihs focuses on the transformation of local patterns into global models illustrated with the transcription of vocal time series into sheet music.

- Katharina Morik discusses the importance of the example representation, because it determines the applicability of methods. For local pattern detection, frequency features are well suited. She shows how to characterize time-stamped data using a frequency model.
- Myra Spiliopoulou gives an overview of local patterns exhibiting temporal structures, namely changes of (learned) concepts.

Seminar Results

Based on the definition of David Hand[2]

data = background model + local patterns + random

seminar participants came up with 12 definitions of what local patterns actually are. These were intensively discussed and we finally agreed on the following:

- Local patterns cover small parts of the data space. If the learning result is considered a function, then global models are a complete function, whereas local patterns are partial.
- Local patterns deviate from the distribution of the population of which they are part. This can be done iteratively — a local pattern can be considered the overall population and deviating parts of it are then determined.
- Local patterns show some internal structure. For example, correlations of features, temporal or spatial ordering attributes, and sequences tie together instances of a local pattern.

Local patterns pose very difficult mining tasks:

- Interestingness measures differ from standard criteria for global models.
- Deviation from background knowledge (global model) requires good estimates of the global mode, where local patterns deviate from the overall distribution.
- Modeling noise (for data cleaning, distinguished from local patterns).
- Automatic feature generation and selection for local patterns (for local patterns other features are more successful than for global models; standard feature selection does not work).
- Internal structures of the patterns (correlations of several features, graphs, sequences, spatial closeness, shapes) can be expressed in several ways, e.g., TCat, constraints.
- Test theory for an extremely large space of possible hypotheses (large sets are less likely, hence global models do not encounter this problem).
- Curse of exponentiality — complexity issues.
- Redundancy of learned patterns.
- Sampling for local patterns speeds up mining and enhances quality of patterns.

[2] David Hand. Pattern Detection and Discovery. In David Hand, Niall Adams and Richard Bolton, editors, *Pattern Detection and Discovery*, Springer, 2002.

– Evaluation: benchmark missing.
– Algorithm issues.

We hope that this books reflects the issues of local pattern detection and inspires more research and applications in this exciting field.

Katharina Morik
Arno Siebes
Jean-Francois Boulicaut