Applications and Research Problems of Subgroup Mining

Willi Klösgen

German National Research Center for Information Technology (GMD)
D-53757 St. Augustin, Germany
kloesgen@gmd.de

Abstract. Knowledge Discovery in Databases (KDD) is a data analysis process which, in contrast to conventional data analysis, automatically generates and evaluates very many hypotheses, deals with complex, i.e. large, high dimensional, multi relational, dynamic, or heterogeneous data, and produces understandable results for those who "own the data". With these objectives, subgroup mining searches for hypotheses that can be supported or confirmed by the given data and that are represented as a specialization of one of three general hypothesis types: deviating subgroups, associations between two subgroups, and partially ordered sets of subgroups where the partial ordering usually relates to time. This paper gives a short introduction into the methods of subgroup mining. Especially the main preprocessing, data mining and postprocessing steps are discussed in more detail for two applications. We conclude with some problems of the current state of the art of subgroup mining.

1 Introduction

Three main application purposes of a data mining task can be distinguished. The predominant purpose is to use mining results to predict or classify future cases, e.g. classify a client or transaction as fraudulent or predict the probability of a group of car drivers causing an accident.

A second purpose of using data mining results is to describe a domain or, more specifically, the dominant dependency structure between variables of the domain, for instance by a summarized overview on the winners of a planned legislation or the client groups whose market shares decreased.

But data are often inconclusive, because not all relevant variables are available, or the description language to describe, for example, subgroups of objects is insufficient. Then the user is satisfied, if some *nuggets* are identified by a mining method. This means that some single hypotheses are validated as interesting by the KDD system. These results are not complete, because they do not allow to predict all new cases or to summarize the whole dependency. For instance, some single subgroups of patients suffering under a special disease are identified without to derive a complete disease profile. Partial results are produced that can give rise to more detailed analyses to be conducted by the domain experts.

These three main application purposes, i.e. classification and prediction, description, and nuggets detection, roughly determine which data mining methods can be applied. For the selection and parametrization of appropriate methods, more special subgoals of applications must additionally be considered which refer to the application dependent properties the discovery results shall have. Such properties relate, for example, to the homogeneity of subgroups, complexity of results, covering of subgroups, (classification) accuracy. Some examples are discussed in sections 4 and 5.

Subgroup mining is a special, but very broadly applicable data mining approach which can be used for these three main analysis goals and many special data mining tasks. It includes several widely used data mining methods such as decision rules and trees, association and sequence rules, change and trend patterns. Subgroup mining searches for subgroups of analysis units (objects or cases: persons, clients, transactions, etc.) that show some type of interesting behavior. The single identified subgroups are local patterns, but can also be seen as a global model when constituting a consistent and non redundant structured set of subgroups. See [1] for a discussion of distinguishing local patterns and global models.

Statistical findings on subgroups belong to the most popular and simple forms of knowledge we encounter in all domains of science, business, or even daily life. We are told such messages as: Unemployment rate is overproportionally high for young men with low educational level, young poor women are infected with AIDS at a much higher rate than their male counterparts, lung cancer mortality rate has considerably increased for women during the last 10 years.

To introduce the subgroup mining paradigm, one has to deal with types of description languages for constructing subgroups, and to define and specialize general pattern classes for subgroup deviations, associations, and sequence patterns. A deviation pattern describes a deviating behavior of a target variable in a subgroup. Deviation patterns rely on statistical tests and thus capture knowledge about a subgroup in form of a verified (alternative) hypothesis on the distribution of a target variable. Search for deviating subgroups is organized in two phases. In a brute force search, alternative search heuristics can be applied to find a first set of deviating subgroups. In a second refinement phase, redundancy elimination operators identify a (best) system of subgroups, usually a subset of the first set or a set including some derived subgroups.

An association pattern identifies two subgroups which have a significant association. Various specializations can be selected to measure the association between two subgroups, e.g. the confidence and support approach of association rules [2]. A sequence pattern identifies a set of subgroups which are partially ordered according to a given ordering type. Thus the partial ordering must satisfy given constraints, e.g. a serial or parallel ordering. The ordering usually relates to time where each object in the database has a time stamp (e.g. errors in a network, transactions of clients). Typically time windows are regarded that contain objects in a special partial order where each object belongs to a special subgroup. Frequent episodes or episode rules [3] are examples of sequence patterns (80 % of windows that contain an error of subgroup B immediately after an error of subgroup A contain later an error of subgroup C). We will concentrate in this paper on specializations of the first general subgroup pattern class, namely deviating subgroups.

Subgroup patterns are local findings identifying subgroups of a population with some unusual, unexpected, or deviating behavior. Thus a single subgroup pattern does not give a complete overview on the population, but refers only to a subset. Nevertheless these local statements (nuggets) are often useful, either because the interest of the analyst is rather modest, or the available data do not allow to derive the complete description of the behavior for the whole population. In a production control application (see section 4), the analyst may be satisfied, if some conditions for the large number of process variables can be identified that lead to a very high quality of the product. Then these conditions will probably be tried to steer a better production process. In a medical application, the available variables can often not describe the dependencies between symptoms and diagnoses, because only a part of the relevant and typically unknown variables is available. But it is useful to know at least some subgroups of patients that can be diagnosed with a high accuracy.

To be useful, subgroup patterns must satisfy at least two conditions. Subgroup descriptions must be interpretable and application relevant, and the reported behavior for the subgroup must be interesting which specifically means that it is statistically significant. This is achieved by the choice of the syntactical type of the description language and background constraints for valid subgroup descriptions, and of a statistical test measuring the statistical significance of a subgroup deviation. An appropriate test is selected according to the type of the analysis question (which is e.g. determined by type of the target variable and the number of populations to be compared in the deviation pattern) and domain dependent preferences of the analyst that e.g. refer to trade-offs such as size of deviation versus size of subgroup. For a detailed discussion of the role of statistical tests in subgroup mining, a classification of specializations of subgroup pattern types, and an overview on search strategies and visualization and navigation operations, see [4].

2 Data and Domain Knowledge for Subgroup Mining

The application domain of a KDD and specifically of a subgroup mining task is represented by data which may have been collected directly for decision support purposes when e.g. a market research company performs a regular survey on a special market and therefore can usually collect the data that are relevant for the analysis (section 5). But the typical data mining application is a secondary data analysis when the data have been generated during an administrative or transaction process (e.g. client database, production data). Secondary data analyses may include problems of quality and appropriateness of given data for the targeted application question. In addition to data, a knowledge base that holds domain knowledge can be exploited in subgroup mining.

The vast collections of already available operational databases are often the impetus for data mining applications. In this case, typically a lot of preprocessing is necessary to combine data from different sources and to adapt data to various analysis problems (see sections 4 and 5). Then data are transformed into structures suitable for efficient access and operation. In table 1, a classification of data types is given which can be applied to select appropriate mining tasks, analysis methods, and data management solutions for an application.

Subgroup mining methods can operate on observation or tansaction data as well as on textual data. When searching for interesting subgroups of documents [5], each document is represented by a set of terms or phrases that are included or relevant for the document, and a subgroup is described by a conjunction of terms. The statistical evaluations of subgroup patterns rely on simple explorative tests, so that some limited representativity of available data can be handled.

Main data type	observation transaction	textual	multimedia	
Representativity	complete population	sample of convenience	random sample	stratified sample
Variable type	binary	categorical	continuous	mixed
Missing data	no	yes		
Conclusiveness	low	medium	high	
Size	moderate	large	very large	vast
Dimensionality	low	medium	high	
Dynamics	static	timely evolving		
Distribution	local	fixed locations	scattered on net	
Object heterogeneity	one object class	multi-valued attributes	multiple object classes	
Time reference	one cross section	two indepen-dent cross sections	series of independent cross sections	longitudinal data
	continuous	time series		
Space reference	point	line	area	surface
Text structure	unstructured	structured parts	hypertext	
Text languages	English	other languages	mixed collection	multilingual text
Hybrid forms	no	mixed obser-vational data	observational & text & multimedia	
Aggregation	micro data	macro data		
Meta data	no	data dictionary	domain knowledge	

Table 1. Dimensions for classifying data and selecting appropriate specializations of subgroup mining patterns

All types of variables can be exploited, using discretization methods for the continuous variables that describe subgroups and different pattern subtypes for categorical, ordinal or continuous target variables. To analyse high dimensional and very large data sets, efficient heuristical search strategies can be selected. For time and space referenced data, special pattern types are used, e.g. for change and trend detection, or by including geographically based clustering methods to derive corresponding selectors for a description. Multi relational description languages can be applied for data with several relations [6].

Subgroup mining methods only exploit some limited form of domain knowledge. Domain knowledge is necessary to achieve a higher degree of autonomy in search and interestingness evaluation of patterns. Current systems typically exploit some meta-data and taxonomical knowledge. Taxonomies are hierarchies on the domains of the variables (e.g. a regional classification). They are necessary for describing interesting results on an appropriate aggregation level. For example, a finding which is based on significant cells in a crosstabulation and holds for nearly all provinces, should be generalized to a higher regional level. Other advantages of taxonomies are listed in table 2.

data reduction	very large datasets, categorical analysis methods
interpretability	avoid nonsense groupings of variable(s) (values)
simplicity	avoid repetitions, find appropriate hierarchical level
ordering	general to specific, find deviations and exceptions
search reduction	pruning of subnodes
greediness	expansion by selectors with higher coverage
focusing	support specification of mining task, select aggregation level

Table 2. Advantages of taxonomies

Another technique to deal with interestingness is based on constraining the search space. Constraints, for example on the combinations of variable values that are potentially useful, can be exploited as domain knowledge to exclude non interesting hypotheses. A typical example is an association rule mining task, when sets of products have to be identified that are often jointly bought by clients in mail-orders. The aim of this analysis may be to evaluate the efficiency of the catalogue issued by the mail-order company. Associations between products that appear on the same page of the catalogue may be defined as uninteresting by a relation between products: same-page(product1, product2). By specifying appropriate constraints, uninteresting hypotheses can be excluded from search. Similar techniques are applied when syntactical biases are defined for the hypotheses language, like a conjunctive description language for subgroups. Then similar variables can be put together in a group, and constraints determine, how many variables of a group and which combinations of variables are selected to build patterns or models.

Interestingness can additionally relate to the *novelty* of the results. The system can exploit former discoveries and analyse the reaction of the user (e.g. rejecting or accepting results by including them into reports).

3 Conditions for Successful Subgroup Mining Applications

Based on different emphasis in using approaches from the statistical, database, machine learning, and visualization fields, various KDD systems have been constructed (see: http://info.gte.com/~kdd/siftware.html) that discover interesting patterns in data. What are now the conditions for a successful application of these tools and when do they provide better results than conventional data analysis methods? Primarily, a data rich application is needed. As a common necessary precondition for every data analysis application, this means at first, that the database is sufficiently conclusive containing problem relevant variables for a lot of approximately representative objects of the domain. Provided that these requirements are satisfied, subgroup approaches prove particularly useful in comparison to the standard statistical methods, when an analysis problem includes a large number of potentially relevant variables, multidimensional relations vary in different subpopulations, no proved (statistical) model was already established for the problem, or surprising results can be expected for subgroups.

For example, in the applications on a production control process (section 4), where settings for production parameters that offer good production results shall be found, we had to study more than 100 variables that may influence the quality of a product. On the other side, if the problem is well understood and there is already a proven model to describe or analyse the domain, then there seems to be little need for new methods. Thus for more than 20 years, a market research application analyses surveys on the financial market (section 5). Banks and insurance companies know their own and their competitors´ client profiles, since they have analysed these survey data with conventional data analysis tools for a long time. We recently supported this application additionally with KDD methods of the *Data Surveyor* system [7]. To detect sudden changes in the current year, KDD methods proved useful in this well studied area too. Unexpected changes in special subgroups (e.g. for subgroups that usually are not studied using only the variables of the main model) can be detected by KDD methods that would have been overlooked in traditional analyses. Moreover, it often helps in a KDD application, if some data analysis experience with the problem is already available.

There are several approaches to combine conventional statistical models with the subgroup mining approach. A first option is to use subgroup mining as a preprocessing and exploration task. Then subgroup mining results are used to familiarize with the data and to get an overview on the relevent influence variables and their form of influence. Based on these subgroup mining results, a model (e.g. regression model) can be specified. An alternative option applies subgroup mining in a postprocessing phase to modeling. The residuals of a statistical model are analysed with subgroup mining methods to detect some non-random residual behavior.

Another special combination option applies probabilistic network techniques to subgroup mining results. To elucidate the dependency structure between found subgroups and the target group, probabilistic nets can be generated for the subgroups which are analysed as derived binary variables.

A further necessary condition for useful KDD applications is, that many disaggregated indices or subgroups must be studied (many dimensions can be relevant for the problem). In a political planning application [8], we analysed the results of socio-economic simulation models to study the consequences of new tax and transfer laws. If the analyst or politician is mainly interested in the overall cost of the new legislation and only in a few fixed groups of winners and losers of the legislation, this can be mostly achieved by studying a small set of cross tabulations. In contrast, a typical subgroup mining application analyses a large set of potentially interesting subgroups for which surprising results can be expected.

4 Application: Production Control

We first discuss a production control application with factory data on process conditions and laboratory data on the quality of products. Process conditions are continuously recorded and quality data several times a day. When fixing a data model, it must be decided how to represent time and space related data. The analysis object must be determined, e.g. a production day. In a simple one-relational model, process variables (pressures, temperatures, etc.) are discretized each hour so that separate variables for time points (each hour) and for different locations in the factory are introduced. Additionally, appropriate derived variables capturing time behaviour like maximum, minimum, average, slope are generated for these time or space indexed variables.

Thus data preparation deals with the selection of a data model for the target data to be mined and the construction of this target data base from the raw data followed by data cleaning and reduction. It is an iterative process including feed back loops, as the next preprocessing steps may generate insights in the domain that require additional data preparations, e.g. by evaluating the appropriateness of derived variables.

Because data from several operational or transaction databases have to be exploited, a separate database is constructed to manage the data for data mining purposes. One advantage of a separate mining database is the better performance. Large scale data mining is very computer time intensive, and many queries must be sent to the data base to calculate the aggregates and cross tabulations. There is also the advantage of no interference with the transaction process and of a well defined state of the data when using a separate mining database. Sometimes only a subset of the raw data is needed in data mining which can be more adequately organized for mining purposes (often relying on an variable oriented or inverted data organization). Disadvantages refer to the time effort that may be necessary for importing data to the separate mining data base, and the storage and management overhead for maintaining two data versions. For some applications, an up to date state of the data may be important.

Decision trees can deal with this high dimensional production data application. Execution time is fast, because of the one step lookahead (hill climbing), non-backtracking search through the space of all possible decision trees. However, decision trees cannot spot interactions among two or more variables and typically there exist problems for highly uneven or skewed distributed dependent variables, e.g. a binary dependent variable with nearly all cases in one of the two classes.

Because of the hill climbing, non backtracking approach, only a limited set of all possible conjunctions of selectors that can be built with the (discretized) values of the variables are examined. Often rule based methods overcome these limitations. Various search strategies exist for rule set generating methods, see [9] for more details. Another popular method generates many trees for the dataset, using statistical relevance tests to prune rules and filter them, and thus selecting best rules out of multiple trees to generate a robust, compact and accurate set of classification rules. This goes back to the observation that each tree has a few good rules and many bad ones. Figure 1 and figure 2 show examples of some rule sets that were derived with the subgroup mining system Explora [9] for this production control application.

The results of figure 1 are very untypical for KDD applications. In this case, all four result properties that are important for a rule set could be achieved: A coverage of the whole target group of bad productions with a small number of three disjoint subgroups that are highly statistically significant (exact 100% rules in this case) could be derived. We have found in the example necessary and sufficient descriptions of the target group. Due to the often noisy data situation and the inconclusiveness of the variables used for the description of target groups, such necessary and sufficient conditions for target groups will usually not be derivable.

```
Problem:    Conditions for extremely bad productions
            target group: QKG1 > 0.25,  QKG3 > 0.61

Pattern:    Probabilistic rules

Strategy:   High accuracy, Disjointness, Recursive
            exhaustive search

Production: Plant A, 1995

            1% of the productions are extremely bad

Disjoint subgroups describing this target group:

  100% of PG1 = 0.7, PG9L4 = 0,       PG2 = 0

  100% of PG8 = 0.3, PG3L8 0.28-0.33

  100% of PG2 = 0.2, QKG2  0.47-0.61, PG3 0.3-0.4,
          PG14  0.4-0.5

These 3 subgroups cover 100% of the target group.
```

Fig. 1. Conditions for very bad products

Figure 2 shows an analysis question with more typical results. These were derived in an exhaustive search process of limited depth which was run recursively. This search strategy can be seen as a generalised beam search, where not only a one level exhaustive expansion of the *n* best nodes is scheduled, but a multi level exhaustive expansion. Thus larger parts of the search space are heuristically processed.

We select a two level expansion procedure. So search in a space of at most two (additional) conjunctions (maximal number of conjunctions selected according to the size of the database, i.e. the number of cases and variable values) is run exhaustively considering however four pruning conditions. A subgroup is not expanded any more, if the statistical significance of the subgroup is already very high, the generality of the subgroup is low (small size), the number of target elements covered by the subgroup is low, or the statistical significance of the subgroup is highly negative. Then a small subset of all found subgroups that exceed a minimum significance level (determined according to the number of statistical tests run in the hypothesis space) is selected that is overlapping below a specified threshold (10% in this case: low overlapping) and covers the target group maximally. This subset selection problem is scheduled as a second search process. Each of the selected subgroups is expanded in a next recursion of this two-phase search and subset selection process with similar prunings and limitations on the number of (additional) conjuncts.

For this application, we use a simple conjunctive description language that does not include any internal disjunctions of values of one variable. Internal disjunctions are often useful for nominal or ordinal variables (section 5). For the predominant continuous description variables of this application, we do not apply a discretization algorithm which is dynamically embedded into the data mining search, but a static generation of taxonomies with hierarchical systems of intervals for the variables in a preprocessing step. Based on visualizations of the variable distributions, appropriate intervals were selected by the domain experts. These static discretizations are much more time efficient than dynamic discretizations, which in this application case and hardware environment (Macintosh) would not have been realizable in an interactive exploration mode with short response times. Some of the drawbacks caused by the fixed static intervals could be mitigated by the hierarchical taxonomical structure.

For a data mining search strategy, in general four components are important. First, the type of the description language has to be selected. Besides simple conjunctive descriptions built with one value selectors (a value is an element of the domain of a variable or an entry in the taxonomy), also negations of one value selectors can be included (marital status ≠ married). A next extension of the complexity of the description language (and search space) allows internal disjunctions of values, such as marital status = married or divorced.

As a second component, a neighborhood operator (expansion operator for general to specific search) generates the neighbors of a description in the partially ordered description space. For a description with internal disjunctions, the neighbors can be constructed by eliminating one of the disjunctive elements in a description.

```
Problem:   Conditions for good productions
           target group: QKG1 < 0.16,  QKG3 < 0.41

Pattern:   Probabilistic rules

Strategy:   High accuracy, Low overlapping, Recursive
            exhaustive search

Production: Plant A, 1995

            20% of the productions are good

Subgroups describing the target group:

       66% of PG2L8 = 0, PG4 <0.18
       94% of PG2L8 = 0, PG4 < 0.18, PG15 > 0.88,
              PG15L8 > 0.88
      100% of PG2L8 = 0, PG4 < 0.18, PG1  = 0.25,
              PG12 0.45–0.55

22% coverage of the target group, 9.8% overlapping

       55% of PG8 = 0.7, PG11L8 0.8-0.85
      100% of PG8 = 0.7, PG11L8 0.8-0.85, PG4 < 0.18,
              PG15L8 > 0.88
      100% of PG8 = 0.7, PG11L8 0.8-0.85, PG1L4 = 0.25

19% coverage of the target group, 5.4% overlapping

       63% of PG1 = 0.5, PG16 0.5-0.6
       90% of PG1 = 0.5, PG16 0.5-0.6, PG15 0.6-0.7,
              PG15L4 0.6-0.7
       86% of PG1 = 0.5, PG16 0.5-0.6, PG8L8 = 0.75,
              PG15L4 0.6-0.7

35% coverage of the target group, 9.4% overlapping

total coverage of target group: 66%
total overlapping: 19%
```

Fig. 2. Conditions for good productions

Third, the quality function measuring the interestingness of a subgroup is important. For these production applications we select a quality function that favours large deviations (and not large subgroups). However, the choice of the quality function cannot only be discussed under domain specific requirements (which in this case request large deviations, i.e. high shares of bad or good productions), but also from search aspects.

Especially, the greediness of search is important. By favouring large subgroups, the greediness of a hill climbing approach can be restricted. Using taxonomies or internal disjunctions restricts the greediness and provides a more patient search [10], so that in each expansion step the size of a subgroup is not restricted too much.

Fourth, the search heuristic is an important further search aspect. Various heuristic strategies for large scale and refinement search are possible. In this application, we have selected the generalized beam search heuristic.

5 Application: Market Analysis

A next exemplary application refers to the analysis of market research data about the financial market´in Germany. All leading financial institutions (banks, insurances) have joined in a common consortium to collect data on the whole market. A market research institute performs surveys by monthly interviewing 2500 persons on their behavior in the market. Several hundreds of variables are collected for each interviewed person. Most of these variables are hierarchically structured, e.g. hierarchical product categories, and have multiple values, e.g. a person can be a client of several banks or use several financial services. Since the data are captured according to a survey plan, each case has an individual weight. Such weighted, multi-valued data provide problems for many data mining systems that only assume simple rectangular one-relational data. This application has been run first with the Explora subgroup mining system and later with Data Surveyor.

Each institution represented in the consortium thus gets survey data on the whole market, included data on the competitors. Therefore the data complement the client databases of each institution. Since the surveys have been regularly collected for some twenty years, the overall database consists of several hundred thousands of cases. It is however an incremental application where the focus is on changes in the current period and in trend detection. Thus the change and trend patterns of subgroup mining are primarily used for the application.

Feature selection techniques are applied to select a subset of relevant independent variables. Sampling methods can select a subset of cases, and missing data treatment clean and amend the raw data. Various techniques to deal with missing data can be used. One possibility is to infer the missing items by exploiting relations with other variables. For example, the gender of a person in a questionnaire could be inferred from other informations (e.g. first name). However, it must be considered whether the fact that a data item is missing (some persons did not fill in their gender in the questionnaire) can provide useful informaton. In another customer retention analysis, the group of clients of a bank that did not fill in this field proved as a client group with a critical tendency that had a high probability to desert to a business competitor.

As the data set is studied for a long time, the overall client profiles are known quite well for the following analytic question: Who does (not) buy product A in the current period (e.g. year)? This question does typically not provide many interesting (novel) answers for any instantiation of A in an institution / product hierarchy. Using the change patterns of subgroup mining [4], one searches for subgroups for which the market share of a product has significantly changed in the current period (month, quarter, year) compared to the previous period. Then quite interesting results are usually detected which belong to two categories.

A first category identifies subgroups which can be explained after some moments of considerations. They are not really novel, since one has found an explanation for the behavior (e.g. some change in legislation, product policy, etc.). But usually one did not think at these subgroups before performing the mining tasks and thus these results are valuable.

A next category identifies subgroups with a behavior change for which one cannot find an easy explanation. These have to be checked in more detail, since they will reveal often errors in the data or some tricky behavior of the interviewers. In one such case, an interviewer could be identified who manipulated the interviews to save his interview time.

Alltogether, these change patterns mainly belong to the nuggets category. Not a whole overview on the change for the current period is derived, but only local subgroups are identified. The change patterns directly identify client subgroups, for which the market share has changed, or in case of trend patterns, show a special type of trend (e.g. linear monotonic increase). This is often more revealing than comparing changes in subgroup patterns that have been statically derived for individual time points. When comparing systems of subgroups (e,g. rule sets derived for several time periods), sensitivity problems of subgroup descriptions usually make the interpretation more difficult.

Good results could also be achieved by combining subgroup mining approaches with conventional modeling techniques. The subgroup mining patterns are inferred for the residuals of a model. Then the model describes the major dependency or trend and the subgroups are identified that deviate from this major line.

Primarily nominal or ordinal variables are applied as explanatory description variables in this application. Thus the type of the description language has been selected to include internal disjunctions. This proved advantageous compared to the simple conjunctive description language, because more general and larger subgroups have been identified which is requested for this application when adjusting the generality-deviation tradeoff. The internal disjunctions further present an overview on the values of a variable that behave similarly (e.g. have an overproportional market share) by including a list of these values and not just the most significant value. In most cases the found internal disjunctions had a clear interpretation. As an example, a list of those German States are identified as an internal disjunction, that constitute the (new) East German States. This disjunction could of course have also been provided as a taxonomical entry in a hierarchy. Internal disjunction languages however contribute to find such groupings dynamically and dependent on a current analytical question. Subsets of values are found that belong together within the context of the current analytic question. These subsets do usually not belong to a fixed taxonomical entry (as the Est German States), but are more loosely connected in the special context. As a special case, missing data is associated to values that play a similar role with respect to the analytic question.

Also for ordinal variables, the internal disjunction approach proved as a competitive alternative to discretization techniques. Thus based on a static prediscretization by a set of intervals, ranges of intervals have been identified that e.g. represent U-form dependencies. For instance in the patient PRIM discretization method [10], such U-form dependencies cannot be detected.

6 Conclusion: Research Problems for Subgroup Mining

The main (data) problems that may occur in a knowledge discovery task are inconclusive data, i.e. the relevant variables are missing, the cases are not representative, there are not enough target cases available, so that the data collected for a primary operational application may not be adequate for the data analysis task. Problems may also exist for redundant data (too many highly correlated data), time and space dependent data, insufficient or inappropriate domain knowledge (no relevant taxonomies to find a good generality level).

The preceding sections have shown that there is already a vast spectrum of subgroup mining variations applied successfully for a broad range of applications. However, KDD is an evolving research area, and some problems have to be solved to provide high quality discovery results.

For the descriptive patterns of subgroup mining, it must be avoided that the descriptions overfit given data. This can mainly achieved with train and test methods which are already successfully applied for classification relying on the evaluation of classification accuracy.

As mentioned in the discussion of figure 2, there are several criteria, e.g. degree of coverage, variance reduction, degree of overlapping, significance, simplicity that are jointly used to assess the quality of a set of subgroups describing a target group. Finding the best set of hypotheses among a large set of significant hypotheses is still a problem. Methods such as tree construction heuristics produce a set of rules arranged as a tree, but usually there may exist better rule sets. One approach for this problem of generating and evaluating a "best" set of hypotheses is discussed in the example of figure 2. Using Bayesian nets for subgroup results is an alternative approach that selects subgroups based on conditional independence.

A next problem area relates to providing adequate description languages. We have already mentioned the limited expressive power of propositional, attributive languages for the production application where additional variables are derived to represent the average or trend in a time or space sequence of measurements. Sometimes *first order language* based approaches can be helpful [11]. Constructive induction is another important approach related to this problem. Here, additional variables are constructed (dynamically during search) that are better suited to describe the given data. Especially for time and space related data, such derived variables can be useful when including descriptive terms based on means, slopes or other (time-) series indicators (see section 4). Dynamically starting a separate search process for finding adequate expressions in a description language is already successfully solved for the problem of discretizing numerical variables and finding geographical clusters of values for regional variables.

Integrating several aspects of interestingness, e.g. statistical significance, novelty, simplicity, usefulness, is a next problem. Future generations of systems will include discovered knowledge in their domain knowledge base to a still higher extent and use these findings for further discovery processes. Thus, they will incorporate more learning and adaptive behavior. Discovery methods could be used to learn from the users by monitoring and analysing their reactions on the discovered and presented findings to assess the novelty facet of interestingness.

For very large datasets including millions of tuples and several hundreds of variables, data reduction is important. Although high performance solutions such as Data Surveyor can extend the applications of KDD methods from the usual boundaries of common KDD systems (10**6 tuples, a few hundreds of fields) by some orders, feature selection and sampling methods are often necessary to provide time efficient interactive discoveries. Interactivity of discovery systems is important because of the explorative nature of the KDD process. Reduction of variables is also important for ensuring clear discovery results.

Explorative data analysis methods often underline the principle of robustness (compare Tukey and other protagonists). This means for KDD that discovery results should not differ too sensitively respective to small alterations of the data, description language or selected values of the dependent variables. As an example, consider the definition of the target group in figure 2. If the specification `QKG3 > 0.61` would be slightly changed to `QKG3 > 0.62` (assuming no large distributional effects for the variable QKG3), this should not lead to a totally diverse set of rules identified in the discovery process. The main concern in KDD has been on accuracy, whereas robustness until now only plays a minor role in discovery research. Interactive visualization techniques can help to analyse different sensitivity aspects [10].

A variety of patterns is available to execute discovery tasks such as identification of interesting subgroups. *Second order discovery* to compare and combine the results for different patterns could be necessary, especially if many analysis questions are issued to the data. An example refers to combining change patterns with static patterns for the two time periods that define the reference of change.

A next point relates to changing data and domain knowledge. This includes incremental mining methods adapting existing results according to the addition or modification of a small number of tuples and comparing new discovery results (for the new data) with the preceding results. Incremental methods operating on sequential (or also parallel) batches of data are also useful for achieving scalability or anytime data mining.

The role of domain knowledge in KDD is restricting search to exclude uninteresting findings and to increase time efficiency. This technique has been mainly applied in *Inductive Logic Programming* approaches, but should be more extensively applied also for the mainstream KDD methods.

Finally there are a lot of technical challenges to ensure efficient and interactive KDD processes. High performance solutions are necessary for VLDB applications. Other problems relate to the integration of KDD systems with other systems such as database systems or statistical packages. One promising development direction is characterized by incorporating KDD functionality within DBMS systems.

References

1. Hand, D.: Data mining - reaching beyond statistics. Journal of Official Statistics 3 (1998).
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, I.: Fast Discovery of Association Rules. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge (1996) 307 – 328.
3. Mannila, H., Toivonen, H., Verkamo, I.: Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery 1 (3) (1997) 259 – 289.
4. Klösgen, W.: Deviation and association patterns for subgroup mining in temporal, spatial, and textual data bases. In: Polkowski, L., Skowron, A. (eds.): Rough Sets and Current Trends in Computing. Lecture Notes in Artificial Intelligence, Vol. 1424. Springer-Verlag, Berlin Heidelberg New York (1998) 1 – 18.
5. Feldman, R., Klösgen, W., Zilberstein, A.: Visualization Techniques to Explore Data Mining Results for Document Collections. In: Heckerman, D., Mannila, H., Pregibon, D. (eds.): Proceedings of Third International Conference on Knowledge Discovery and Data Mining (KDD-97). AAAI Press, Menlo Park (1997).
6. Wrobel, S.: An Algorithm for Multi-relational Discovery of Subgroups. In: Komorowski, J., Zytkow, J. (eds): Principles of Data Mining and Knowledge Discovery. Lecture Notes in Artificial Intelligence, Vol. 1263. Springer-Verlag, Berlin Heidelberg New York (1997) 78 – 87.
7. Siebes, A.: Data Surveying: Foundations of an Inductive Query Language. In: Fayyad, U., Uthurusamy, R. (eds.): Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDDM95). AAAI Press, Menlo Park, CA: (1995).
8. Klösgen, W.: Exploration of Simulation Experiments by Discovery. In: Fayyad, U., Uthurusamy, R. (eds.): Proceedings of AAAI-94 Workshop on Knowledge Discovery in Databases. AAAI Press, Menlo Park (1994).
9. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA (1996).
10. Friedman, J., Fisher, N.: Bump Hunting in High-Dimensional Data. Statistics and Computing (1998).
11. Quinlan, R.: Learning Logical Definitions from Relations. Machine Learning 5(3) (1990).