# Profiling Examiners using Intelligent Subgroup Mining

**Martin Atzmueller** and **Frank Puppe**
Department of Computer Science
University of Würzburg
97074 Würzburg, Germany
{atzmueller, puppe}@informatik.uni-wuerzburg.de

**Hans-Peter Buscher**
DRK-Kliniken Berlin-Köpenick,
Clinic for Internal Medicine II,
12559 Berlin, Germany
buscher.dhp@t-online.de

## Abstract

The demand for effective knowledge discovery methods in a clinical setting is growing: the number of hospital information systems and medical documentation systems in routine-use increases rapidly. Then, often high-quality collections of electronic patient records are available for statistical analysis. One interesting issue concerns the quality of the examinations records which depends both on the examination quality and the documentation habits of the individual examiners. We apply a subgroup mining approach for explorative and descriptive data mining to tackle this issue, and we provide a case study of the proposed approach using data from a fielded system in the medical domain.

Purely automatic data mining methods often suffer from the limitation that too many uninteresting results are presented to the user. In order to improve upon this situation, we propose two strategies: we use background knowledge, if available, and provide suitable visualizations for guiding the discovery process. The context of the presented approach is a knowledge-based documentation and consultation system.

## 1 Introduction

The available data in clinical settings is growing with a rapid pace. More and more hospitals use medical information systems and/or (knowledge-based) documentation systems that enable the storage of electronic patient records (EPRs). Then, subsequent analysis of high-quality EPRs is a promising option. The quality of the stored examination records is determined by the documentation habits of the examiners, i.e., depending on the experience and training of the individual examiners. Therefore, the identification and analysis of documentation patterns of different examiners is a crucial task to improve the quality of the examinations and therefore of the whole database of patient records.

We propose a subgroup mining approach to analyze the inter-individual documentation quality of the examiners. Subgroup mining or subgroup discovery [Wrobel, 1997; Klösgen, 2002] is a promising technique for explorative and descriptive medical data mining that aims to discover ”interesting” subgroups of individuals. Then, the subgroups can be defined as a subset of the target population with a distributional unusualness concerning a certain property we are interested in, e.g., in the subgroup of smokers with a positive family history the risk of coronary heart disease is significantly higher than in the general population.

Subgroup mining is especially suited for the sketched analysis task in the medical domain, since it does not necessarily focus on finding complete relations between the specific target concept and the explaining variables; instead, interesting partial relations are sufficient. Due to this criterion the discovered patterns do not necessarily fulfill high support criteria, which are necessary for other prominent data mining approaches, e.g., methods for association rule discovery [Agrawal and Srikant, 1994]. Furthermore, subgroup discovery methods do not depend on support measures, but on a quality function which is flexibly defined according to the criteria of the user.

Usually the ultimate goal of knowledge discovery methods is to identify novel, potentially useful, and interesting knowledge. However, in real-world settings novelty and interestingness criteria of the user often cannot be fully satisfied: quite similar to a search query submitted to a web search engine, (e.g., Google), the application of purely automatic methods can yield a huge number of (possibly uninteresting) results which are hard to handle. Then, a ’query refinement’ needs to be considered. In order to perform the discovery process more intelligently, we propose the combination of a semi-automatic subgroup mining method guided by visualization and background knowledge.

We exemplify the approach in a case study based on the knowledge-based documentation and consultation system for sonography SONOCONSULT [Huettig *et al.*, 2004], which is in routine use in the DRK-hospital in Berlin/Köpenick: we identify profiles of examiners concerning their documentation habits for general quality control and management.

The rest of the paper is organized as follows: In Section 2 we introduce our method, i.e., a process model for knowledge-intensive subgroup mining. We describe suitable background knowledge for integration into the mining method and a visualization method to guide the user in the interactive discovery process. Finally, we provide the results of a case study of the presented approach with a fielded system in the medical domain in Section 3. We conclude with a summary of the paper in Section 4.

## 2 Methods: The Semi-Automatic Process for Knowledge-Intensive Subgroup Mining

Subgroup mining aims to discover "interesting" subgroups of individuals that are described by relations between independent (explaining) variables and a dependent (target) variable, rated by a certain interestingness measure. For example, two possible criteria are the difference in the distribution of the target variable concerning the subgroup and the general population, and the subgroup size. Subgroup mining does not necessarily focus on finding complete relations; instead partial relations, i.e., (small) subgroups with "interesting" characteristics can be sufficient.

In this section we first describe the process model for intelligent subgroup mining. After that, we define the subgroup mining task, and discuss the elements of the proposed process model in detail, i.e., helpful background knowledge applied for subgroup mining, and the core visualization method to guide the subgroup mining process. Finally, we discuss related work.

### 2.1 Process Model

The general goal of a subgroup mining task is to identify a set of highly interesting, diverse subgroups. Both the quality measures for the subgroup and the redundancy criteria heavily depend on the goals of the user. A purely automatic approach is often appropriate, if the analysis goals of the user are fixed during the search process. However, if the user wants to test specific hypotheses or already has a lot of background knowledge and experiences about the analysis domain, then an automatic search method may not always be transparent enough.

In the proposed mining process both interactive and automatic elements are combined: the automatic methods can be used to identify useful starting points for analysis, or for a quick "what if" analysis of the current situation. The presented approach includes the background knowledge and experiences of the user in order to focus the mining method on the interesting patterns, and to restrict the search space. Then, direct user interaction enables an *active mining* approach (e.g., [Gamberger *et al.*, 2003]). In this approach, the user is directly integrated into the subgroup discovery process and can manipulate the subgroup descriptions interactively. The process model is depicted in Figure 1.

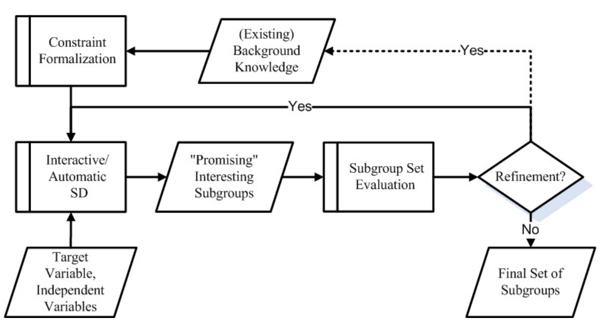


Figure 1: The Knowledge-Intensive Semi-Automatic Subgroup Mining Process

### 2.2 Subgroup Mining

We first introduce our knowledge representation schema before defining the subgroup mining task. After that, we describe the background knowledge and the visualization method used in the proposed subgroup mining process.

**General Definitions**

Let $\Omega_A$ the set of all attributes. For each attribute $a \in \Omega_A$ a range $dom(a)$ of values is defined. Furthermore, we assume $\mathcal{V}_A$ to be the (universal) set of attribute values of the form $(a : v)$, where $a \in \Omega_A$ is an attribute and $v \in dom(a)$ is an assignable value. A diagnosis attribute is represented by a binary attribute, i.e., for a diagnosis attribute $d \in \Omega_D, \Omega_D \subseteq \Omega_A$ we define a (boolean) range $dom(d) = \{established, not\ established\}$. Let $CB$ be the case base containing all available cases. A case $c \in CB$ is defined as a tuple $c = (\mathcal{V}_c, \mathcal{D}_c)$, where $\mathcal{V}_c \subseteq \mathcal{V}_A$ is the set of attribute values observed in the case $c$. The set $\mathcal{D}_c \subseteq \mathcal{V}_A$ is the set of diagnoses describing the *solution* of this case.

**Basic Subgroup Mining** A subgroup mining task mainly relies on the following four main properties: the target variable, the subgroup description language, the quality function, and the search strategy. The target variable may be binary, nominal or numeric. Depending on its type, there are different analytic questions, e.g., for a numeric target variable we can search for significant deviations of the mean of the target variable.

A subgroup mining problem encapsulates the target variable, the search space of independent variables, the general population, and additional constraints.

**Definition 1 (Subgroup Mining Problem).** *A subgroup mining problem $SP$ is defined as the tuple*

$$SP = (T, A, C, CB),$$

*where $T \in \Omega_A \cup \mathcal{V}_A$ is a target variable. $A \subseteq \Omega_A$ is the set of attributes to be included in the subgroup discovery process. $CB$ is the case base representing the general population used for subgroup mining. $C$ specifies (optional) constraints for the discovery method. We define $\Omega_{SP}$ as the set of all possible subgroup mining problems.*

The definition above allows for arbitrary target variables. However, for our analytic questions we will focus on binary target variables, i.e., $T \in \mathcal{V}_A$.
The description language specifies the individuals from the reference population belonging to the subgroup.

**Definition 2 (Subgroup Description).** *A subgroup description $sd = \{e_i\}$ consists of a set of selection expressions (selectors) $e_i = (a_i, V_i)$ which are selections on domains of attributes, i.e., $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. A subgroup description is defined as the conjunction of its contained selection expressions. We define $\Omega_{sd}$ as the set of all possible subgroup descriptions.*

A quality function measures the interestingness of the subgroup (c.f., [Klösgen, 2002] for examples).

**Definition 3 (Quality Function).** *A quality function*

$$q : \Omega_{sd} \times \Omega_{SP} \rightarrow R$$

*evaluates a subgroup description $sd \in \Omega_{sd}$ given a subgroup mining problem $SP \in \Omega_{SP}$. It is used by the search method to rank the discovered subgroups during search.*

For binary target variables, examples for quality functions are given by

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N - n}}\,, \quad q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)}\,,$$

where $p$ is the relative frequency of the target variable in the subgroup, $p_0$ is the relative frequency of the target variable in the total population, $N = |CB|$ is the size of the total population, and $n$ denotes the size of the subgroup. In contrast to the quality function $q_{BT}$ (**B**inomial **T**est), the quality function $q_{RG}$ (**R**elative **G**ain) only compares the target shares of the subgroup and the total population measuring the relative gain. Therefore, a suitable support thresholds is necessary to discover significant subgroups.

An efficient subgroup search strategy is necessary, since the search space is exponential concerning all the possible selectors of a subgroup description: commonly, a beam search strategy is used because of its efficiency [Klösgen, 2002]. We apply a modified beam search method, where an initial subgroup description can be selected as the initial value for the beam. Beam search iteratively expands the $k$ best subgroup descriptions by adding the selector that provides the best quality improvement. Iteration stops, if the quality as evaluated by the quality function $q$ does not improve any further.

For the characterization of the discovered subgroups we have two alternatives: Besides the principal factors contained in the subgroup description there are also supporting factors. These are attribute values $supp \subseteq \mathcal{V}_A$, which are characteristic for the containing subgroup, i.e., the value distributions of their corresponding attributes (supporting attributes) differ significantly comparing two populations: the true positive cases contained in the subgroup and non-target class cases contained in the total population. In addition to the principal factors the supporting factors can also be used to statistically characterize a discovered subgroup, as described, e.g. in [Gamberger and Lavrac, 2002].

**Background Knowledge for Subgroup Mining**

There are different classes of background knowledge which can be used in the knowledge-intensive process for subgroup mining, e.g., constraints, ontological knowledge, and abstraction knowledge. Knowledge acquisition is always expensive, so its costs should be minimized. Sometimes knowledge can be derived from already formalized knowledge, e.g., we can derive constraints from ontological knowledge, and thus reduce its acquisition costs. In the following, we summarize the individual knowledge elements; we refer to [Atzmueller *et al.*, 2005] for a more detailed discussion.

Constraints restrict the search process/space by specifying the attributes and attribute values of interest. In addition, a set of attribute values can be used to define additional meta values specific to the application domain. For example, for the diagnosis *cirrhosis of the liver* the values *possible* and *probable* can be defined as a disjunctive attribute value. Furthermore, constraints can also include quality and syntactical constraints that filter the mined patterns during the discovery process.

Ontological knowledge includes information about the domain ontology, e.g., abnormality information/normality information about attribute values indicating either abnormal/pathological states, or the normal state. For example, consider the attribute temperature with the value range $dom(\text{temperature}) = \{\text{normal}, \text{marginal}, \text{high}, \text{very high}\}$. The values *normal* and *marginal* denote normal states of the attribute, while the values *high* and *very high* describe abnormal states. Using abnormality information, we can define meta values containing several attribute values with certain abnormality categories.

Similarity information about attribute values relates to the relative similarity between attribute values. Significant similarities between attribute values can indicate that the respective values can be combined into a new value. Then, appropriate meta values need to be defined. A high attribute weight specifies, that an attribute is relatively important.

Ordinality information is used to indicate the ordinal attributes which can be used to construct certain 'ordinal groups', e.g., summarizing certain consecutive age groups. In general, specifying appropriate meta values can significantly increase the interpretability of mined subgroup patterns for the domain specialist (c.f. Section 3).

Derived attributes (abstraction knowledge) play a special role in the mining process. These attributes are constructed according to the needs of the user, e.g., intermediate concepts which are not contained in the set of basic attributes can be modelled, or attributes can be constructed such that missing values are minimized.

In Table 1, we summarize the different classes and types of background knowledge (CK = constraint knowledge, OK = ontological knowledge, AK = abstraction knowledge). We show their characteristics in terms of the 'derivable knowledge' if applicable, their costs, and their potential contribution to restricting the search space and/or focusing the search process for a qualitative comparison. The individual ratings are based upon our experiences and feedback provided by the domain specialists, e.g., during the case study in Section 3. Considering the costs/impact of the knowledge elements for subgroup mining, the label - indicates no cost/impact; the labels +, ++, and +++ indicate increasing costs and impact. A +(+) signifies, that the respective element has low costs if it can be derived/learned, and moderate costs otherwise. Similarly ++(+) indicates this for moderate and high costs, respectively.

| *Knowledge* | | *Derivable* | *Cost* | Search Space | |
|---|---|---|---|---|---|
| *Class* | *Type* | *Knowledge* | | Restr. | Focus |
| CK | Syntactical Constr. | – | + | + | + |
| CK | **Quality Constr.** | – | + | ++ | ++ |
| CK | Attr. Values Constr. | – | +(+) | + | + |
| CK | Meta Values Constr. | – | +(+) | – | ++ |
| CK | **Attributes Constr.** | – | +(+) | ++ | ++ |
| OK | **Normality** Info | Attr. Val. Constr. | + | + | ++ |
| OK | Abnormality Info | Attr. Val. Constr. | ++ | + | ++ |
| | | Meta Val. Constr. | | – | ++ |
| OK | Similarity Info | Meta Val. Constr. | ++(+) | – | ++ |
| OK | **Ordinality Info** | Meta Val. Constr. | + | ++ | +++ |
| OK | Attr. Weights | Attr. Constraints | +(+) | + | ++ |
| AK | ***Derived Attributes*** | Derived Attributes | +++ | +++ | +++ |

Table 1: Background Knowledge for Subgroup Mining

The most important types of background knowledge with an especially good cost/benefit ratio concerning the subgroup mining task are indicated in bold type.

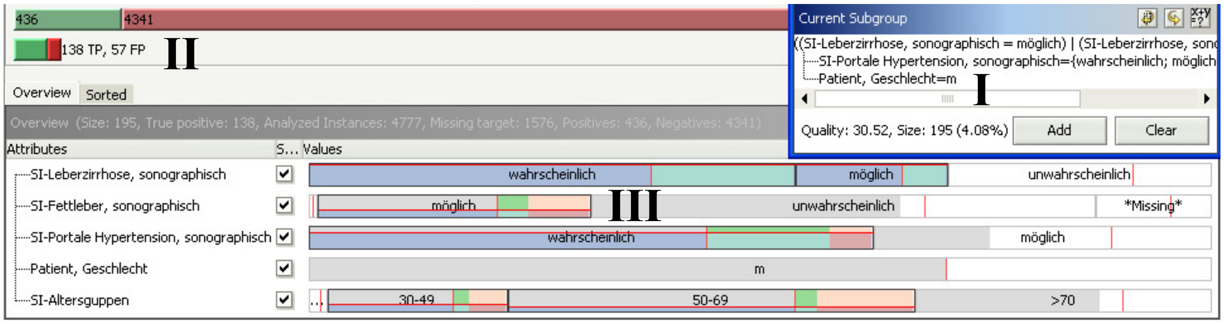


Figure 2: The Zoomtable

**Guiding Subgroup Mining by Visualization Techniques** In this section we present the main visualization for subgroup discovery, i.e., the *zoomtable* depicted in Figure 2. This visualization is associated with the *current subgroup view* (Annotation I) showing the target variable and the selectors of the current subgroup. The bars (Annotation II) depict the target distributions in the whole population (upper bar), and in the subgroup. The left part of a bar shows the positives, the right part the negative instances. The zoomtable (Annotation III) shows the distribution of the data restricted to the currently selected subgroup. Each row of the zoomtable shows the value distribution of a specific attribute. The width of a cell relates to the frequency of an attribute value. The zoomtable is updated when the user modifies the current subgroup, e.g., by adding a selector from the zoomtable.

Figure 3 shows a row of the zoomtable concerning a binary attribute with the values *yes* and *no*. The important parameters for subgroup mining w.r.t. a "future" subgroup are the *subgroup size* – given by the width of a specific selector cell, and the *target share* (precision), i.e., the share of subgroup instances containing the target variable (positive instances). In the current subgroup $SG_c$, (a) indicates the

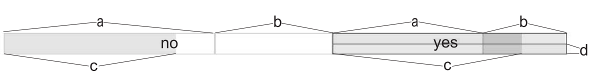


Figure 3: The Zoomtable – Detail View

(currently) positive instances, and (b) denotes the negative ones. In the 'next' subgroup $SG_n$, i.e., including the particular attribute value, (c) shows the positive instances for this subgroup, which can be compared to (a). So, if (c) is larger than (a), then the precision increases adding this selector. Finally, (d) shows the gain in precision, comparing the subgroups $SG_c$ and $SG_n$: if the height of (d) is zero, the precision does not increase. If it fills the entire bar, then the precision reaches 100%.

The zoomtable enables the user to directly manipulate the subgroup and to estimate the effects of individual selectors. Furthermore, interesting attributes and their values are easy to spot due to the visual markers in the respective cells. Then, an active subgroup mining approach (c.f., [Gamberger *et al.*, 2003]) can be implemented quite easily.

## 2.3 Related Work

The application of subgroup mining especially for the medical domain using the guidance of an expert is described in [Gamberger and Lavrac, 2002; Gamberger *et al.*, 2003]. This active approach stresses the interaction between the expert and the system to identify interesting subgroups. However, in the semi-automatic process mainly the parameters of the search process can be adapted. In our semi-automatic process, the domain specialist can adapt the subgroup mining problem by including background knowledge, and modifying the search process directly guided by interactive visualizations.

The proposed interactive core component, i.e., the zoomtable visualization was inspired by the *InfoZoom* system [Spenke, 2001]. InfoZoom also visualizes the value distributions of attributes in single rows of a table, and also allows the user to zoom in on individual values. However, our approach extends this idea significantly, since we also guide the user during the subgroup mining process by visualizing additional quality parameters directly in the zoomtable, e.g., the future target share or the gain of a specific selector. Changes in the zoomtable, e.g., adding/removing selectors to the current subgroup (description) are also visualized dynamically.

Using background knowledge to constrain the search space and pruning hypotheses during the search process has been proposed in ILP approaches. [Weber, 2000] proposes *require*- and *exclude*-constraints for attribute – value pairs, in order to prune the search space. [Zelezny *et al.*, 2003] integrate constraints into an ILP approach as well; the used constraints are mainly concerned with syntactical and quality constraints w.r.t. the discovered subgroups.

The main difference between the presented approach and the existing approaches is the fact, that we are able to integrate several new types of additional background knowledge. This knowledge can be refined incrementally according to the requirements of the mining task. In our process model for semi-automatic and knowledge-intensive subgroup mining we aim to focus the discovery method on the interesting patterns using background knowledge. Then, interactive exploration is made more convenient, since mostly interesting patterns/factors are presented. Furthermore, we apply a novel visualization technique in an active and user-centered approach that is usually more transparent for (experienced) users.

## 3 Results: Case Study

In this section we describe a case study for the application of the proposed subgroup mining process.

We first introduce the analysis task w.r.t. its clinical relevance. Then, we describe the documentation and consultation system SONOCONSULT. After that, we present and discuss the results of the case study.

### 3.1 Profiling Examiners for Quality Control

Our application domain is the domain of sonography. Sonographic examination and documentation is highly dependent on the skills of the examiners. Individual examiners rotate according to a defined schedule (e.g., every 6 months). Before performing the examinations, they get special training and can always consult experienced colleagues. However, while performing the examination they are on their own. Then, it is easy to see that the quality of the examinations is dependent on the individual experience and skills of the examiners. Therefore, documentation and interpretation habits of examiners may differ significantly, which is problematic considering the consistency and quality of the documented examinations; e.g., some examiners may be more competent in identifying specific symptoms concerning certain diagnoses or organ systems than others.

While a gold standard for the correct examination and documentation is not available in sonography, the detection of systematic discrepancies among different examiners is clinically important in itself. To identify deviations regarding the documentation habits of examiners, subgroup mining is used to discover novel and unexpected (documentation) patterns, i.e., certain symptom combinations that are observed significantly more (in-)frequently in conjunction with certain examiners.

### 3.2 The Documentation and Consultation System SonoConsult

We use cases taken from the SONOCONSULT system [Huettig *et al.*, 2004] – a medical documentation and consultation system for sonography – which has been developed with the knowledge system D3 [Puppe, 1998]. The system is in routine use in the DRK-hospital in Berlin/Köpenick and documents an average of about 300 cases per month. These are detailed descriptions of findings of the examination(s), together with the inferred diagnoses (binary attributes). The derived diagnoses are usually correct as shown in a medical evaluation (c.f. [Huettig *et al.*, 2004]), resulting in a high-quality case base with detailed case descriptions. The applied SONOCONSULT case base contains $7096$ cases. The domain ontology contains 427 basic attributes with about 5 symbolic values on average, 133 symptom interpretations, which are rule-based abstractions of the basic attributes, and 221 diagnoses.

### 3.3 Results

The domain specialist performed subgroup mining considering individual diagnostic areas and organ systems, e.g., liver and kidney diseases, using the [VIKAMINE, 2005] system. Then, the relevant factors that were important for deriving the diagnoses of a certain area were identified; these were then provided to the subgroup method in order to constrain the search space and to focus the search method. Furthermore, the domain specialist provided normality information to filter out some uninteresting *normal* values, e.g., *liver vessels = normal*. Meta Values were defined to build disjunctive meta values, e.g., *liver plasticity = moderately or strongly reduced*. Additionally, several derived attributes (abstractions) were defined to limit missing values. For example, diagnostic attributes like *cirrhosis of the liver* were either defined or tuned in order to minimize missing values by providing a default *normal* value. After that, the proposed process model was applied, using beam-search as the automatic component for subgroup mining.

We show examples of the results in Table 2, considering liver diseases, especially focussing on *cirrhosis of the liver*. The cases that were used in the case study were acquired by 8 different examiners (E1 - E8). Concerning liver examinations, each examiner contributed 200-600 cases, resulting in a total population of 3931 cases where an examination of the liver was performed. Then, we analyzed the individual factors concerning the individual examiners as the target variable (column *E*). We used the relative gain quality function $q_{RG}$ (c.f., Section 2.2), which was easy to interpret for the experts. Then, deviations concerning findings or combinations of findings were measured.

Each row of the table depicts a subgroup with the subgroup parameters *Size* (subgroup size), *TP* (true positives), *FP* (false positives), *Pop.* (the defined population), the default and subgroup target share $p_0$ and $p$, respectively, and *RG*, i.e., the value of the relative gain quality function $q_{RG}$.

| # | E | LP | | LS | | LE | | LV | | LC | | Subgroup Parameters | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mr | sr | uk | kn | mi | si | rp | tp | po | pr | Size | TP | FP | Pop. | p0 | p | RG |
| 1 | E1 | | | X | | | | | | | | 221 | 44 | 177 | 2295 | 0.164 | 0.199 | 0.24 |
| 2 | E1 | | | | | | X | | | | | 435 | 41 | 394 | 2295 | 0.164 | 0.094 | -0.51 |
| 3 | E1 | X | X | | | | | | | | | 420 | 28 | 392 | 2295 | 0.164 | 0.066 | -0.71 |
| 4 | E1 | | | | X | | | | | | | 13 | 0 | 13 | 2295 | 0.164 | 0 | -1.19 |
| 5 | E2 | | X | | | | | | | | | 248 | 19 | 229 | 2295 | 0.123 | 0.076 | -0.43 |
| 6 | E2 | | | | | | | X | | | | 689 | 25 | 664 | 2294 | 0.123 | 0.036 | -0.8 |
| 7 | E3 | X | X | | | | | X | | X | | 129 | 91 | 38 | 2294 | 0.129 | 0.705 | 5.12 |
| 8 | E3 | X | | | | | | | | | | 248 | 116 | 132 | 2295 | 0.128 | 0.467 | 3.01 |
| 9 | E3 | | | | | | | X | | X | X | 385 | 131 | 254 | 2294 | 0.129 | 0.34 | 1.87 |
| 10 | E3 | X | X | | | | | | | | | 420 | 132 | 288 | 2295 | 0.128 | 0.314 | 1.64 |
| 11 | E3 | | | | X | | | | | | | 13 | 4 | 9 | 2295 | 0.128 | 0.307 | 1.59 |
| 12 | E3 | | | | | | | | X | | X | 102 | 0 | 102 | 2294 | 0.13 | 0 | -1.14 |
| 13 | E5 | | | | X | | | | | | | 13 | 9 | 4 | 2295 | 0.057 | 0.692 | 11.8 |
| 14 | E5 | X | | | | X | X | | | | | 227 | 85 | 142 | 2295 | 0.057 | 0.374 | 5.89 |
| 15 | E5 | X | | | | | | | | | | 248 | 87 | 161 | 2295 | 0.057 | 0.35 | 5.45 |
| 16 | E5 | X | X | | | | | | | | | 420 | 96 | 324 | 2295 | 0.057 | 0.228 | 3.18 |
| 17 | E5 | | | | | | | | | X | X | 440 | 56 | 384 | 3918 | 0.053 | 0.127 | 1.46 |
| 18 | E5 | X | X | | | X | X | X | | X | X | 271 | 39 | 232 | 2294 | 0.057 | 0.143 | 1.61 |
| 19 | E5 | | | X | | | | | | | | 221 | 6 | 215 | 2295 | 0.057 | 0.027 | -0.55 |
| 20 | E5 | | | X | | X | | X | | X | X | 109 | 0 | 109 | 2294 | 0.058 | 0 | -1.06 |

**LP = Liver Plasticity**
mr = moderately reduced
sr = strongly reduced

**LS = Liver surface**
uk = uneven, knotty
kn = knaggy

**LC = Cirrhosis of the liver**
po = possible
pr = probable

**LE = Liver Echogenicity**
mi = moderately increased
si = strongly increased

**LV = Liver Vessels**
rp = rarefication of portal branches
tp = tapering of portal branches

Table 2: Interesting subgroups and individual factors concerning liver diseases. The first line depicts the subgroup (target variable *Examiner=E1*) described by *Liver surface = uneven, knotty* with a target share of 19.9% ($p$) in the subgroup compared to 16.4% ($p_0$) in the total population with a relative gain of 24% (RG).

Applying the process model, the domain specialist considered the visualization component very helpful, since it enabled an easy step by step analysis: single factors could be identified first, and then subgroups were refined. Furthermore, subgroups discovered by the automatic search method were also validated and refined interactively.

### 3.4 Discussion

The results in Table 2 show significant differences in the documentation habits of the individual examiners. Negative relative gain (RG) values indicate that the examiner documented/interpreted certain findings less frequently than his colleagues, while a positive relative gain indicates the opposite. For a comprehensive overview, we also show some single factors in addition to significant combinations, which were also very interesting for the domain specialist. Especially significant deviations are shown in lines 7, 14 and 15, which are very descriptive for the respective examiners. Line 7 also shows a significant correlation with the diagnosis *cirrhosis of the liver* combined with the relevant findings.

Lines 4, 11, and 13 show a surprising result: the examiners E3 and E5 are the only examiners that document a specific finding, i.e., *Liver surface = knaggy* in comparison to their colleagues. Further investigation turned up that the specific attribute value was added to the consultation system in a later step. Therefore, only some examiners had the opportunity to use this finding.

Furthermore, as shown in the table, examiner E5 (lines 14-20) deserves special attention, since the shown documentation habits differed most significantly compared to the peer examiners. Especially interesting were the subgroups depicted in line 17, 18 and 20: it is easy to see that examiner E5 documents a *cirrhosis of the liver = probable or possible* more frequently than his peers. An even more significant subgroup is shown in line 18 that shows a specialization of the subgroup in line 17. For the very indicative finding combination in line 20 (regarding the diagnosis cirrhosis of the liver) even no case of E5 could be identified. It is striking that E5 uses very special patterns for inferring the diagnosis *cirrhosis of the liver* compared to his colleagues: e.g., symptoms of plasticity are much more frequent (lines 14-16) whereas *liver surface = uneven, knotty* is significantly infrequent (lines 19, 20).

In summary, these results show a high variability of documentation and interpretation habits of the different examiners. They indicate the need for further prospective studies. These results are a starting point for initiating a discussion on training or standardization actions to increase the inter-examiner homogeneity of the sonographic reports.

## 4 Conclusion and Outlook

In this paper we presented an approach for semi-automatic and knowledge-intensive subgroup mining. We exemplified the approach in a case study in the medical domain of sonography, where we were able to extract interesting profiles of examiners concerning their documentation habits. The proposed approach applies background knowledge and visualization to guide the subgroup mining process, which was regarded as extremely important by the domain specialist. The obtained results are a first step toward surveying the documentation performance of individual examiners, and to support their learning phase.

In the future, we are planning to embed a component for subgroup analysis in knowledge-based documentation systems directly. A prerequisite is a comprehensive analysis applying the presented method to identify interesting patterns. Then, using these patterns, the completeness of findings regarding specific examiners can be checked instantly. This provides a transparent survey of general documentation habits and the potential for training certain examiners.

## References


[Agrawal and Srikant, 1994] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[Atzmueller *et al.*, 2005] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In *Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI-05), to appear*, 2005.

[Gamberger and Lavrac, 2002] Dragan Gamberger and Nada Lavrac. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.

[Gamberger *et al.*, 2003] Dragan Gamberger, Nada Lavrac, and Goran Krstacic. Active Subgroup Mining: a Case Study in Coronary Heart Disease Risk Group Detection. *Artificial Intelligence in Medicine*, 28:27–57, 2003.

[Huettig *et al.*, 2004] Matthias Huettig, Georg Buscher, Thomas Menzel, Wolfgang Scheppach, Frank Puppe, and Hans-Peter Buscher. A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography. *Medizinische Klinik*, 99(3):117–122, 2004.

[Klösgen, 2002] Willi Klösgen. *Handbook of Data Mining and Knowledge Discovery*, chapter 16.3: Subgroup Discovery. Oxford University Press, New York, 2002.

[Puppe, 1998] Frank Puppe. Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-Kit D3. *Intl. Journal of Human-Computer Studies*, 49:627–649, 1998.

[Spenke, 2001] Michael Spenke. Visualization and Interactive Analysis of Blood Parameters with InfoZoom. *Artificial Intelligence In Medicine*, 22(2):159–172, 2001.

[VIKAMINE, 2005] http://ki.informatik.uni-wuerzburg.de/projects/dm, 2005.

[Weber, 2000] Irene Weber. Levelwise search and Pruning Strategies for First-Order Hypothesis Spaces. *Journal of Intelligent Information Systems*, 14:217–239, 2000.

[Wrobel, 1997] Stefan Wrobel. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. 1st European Symposion on Principles of Data Mining and Knowledge Discovery*, pages 78–87, Berlin, 1997. Springer Verlag.

[Zelezny *et al.*, 2003] Filip Zelezny, Nada Lavrac, and Saso Dzeroski. Using Constraints in Relational Subgroup Discovery. In *International Conference on Methodology and Statistics*, pages 78–81. University of Ljubljana, 2003.