Journal of Universal Computer Science, vol. 11, no. 11 (2005), 1752-1765
submitted: 1/9/05, accepted: 1/10/05, appeared: 28/11/05 © *J.UCS*

Semi-Automatic Visual Subgroup Mining using VIKAMINE

Martin Atzmueller
(University of Würzburg, Germany
atzmueller@informatik.uni-wuerzburg.de)

Frank Puppe
(University of Würzburg, Germany
puppe@informatik.uni-wuerzburg.de)

Abstract: Visual mining methods enable the direct integration of the user to overcome major problems of automatic data mining methods, e.g., the presentation of uninteresting results, lack of acceptance of the discovered findings, or limited confidence in these. We present a novel subgroup mining approach for explorative and descriptive data mining implemented in the VIKAMINE system. We propose several integrated visualization methods to support subgroup mining. Furthermore, we describe three case studies using data from fielded systems in the medical domain.
Key Words: subgroup mining, visualization, data analysis, data mining
Category: I.2.1, I.2.6, H.5.1, H.5.2

1 Introduction

Knowledge discovery in databases (KDD) tasks [Fayyad *et al.*, 1996] aim 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 (e.g., [Atzmueller *et al.*, 2005b]): quite similar to a search query submitted to a web search engine, the application of purely automatic methods can yield a huge number of results that are possibly uninteresting, not novel or not useful. Furthermore, often automatic methods are not transparent enough for the user. Visual mining approaches (e.g., [Wills and Keim, 2002; Klösgen and Lauer, 2002]) are often a promising option since visual techniques can increase the effectiveness of automatic methods by utilizing the perception and general knowledge of the user. Furthermore, user integration can also increase the degree of confidence concerning the discovered findings. In an active mining approach [Motoda, 2002; Gamberger *et al.*, 2003] automatic methods can be supplemented by visual methods effectively.

We apply a subgroup mining approach (e.g., [Wrobel, 1997; Klösgen, 2002]) for explorative and descriptive data mining to obtain an overview of the dependencies between a specific target (dependent) variable and usually many explaining (independent) variables. A subgroup can be defined as a subset of the target population with a (distributional) unusualness concerning a certain property we are interested in; e.g., the risk of coronary heart disease (target variable) is significantly higher in the subgroup of smokers with a positive family history than in the general population. The unusualness (or interestingness) of a subgroup is formalized by a user-defined quality function.

In general, there are three fundamental paradigms for exploratory data discovery systems: search, visualization, and interactive navigation. We combine automatic and interactive search methods, and present a novel visualization-based approach for active subgroup mining supporting exploration and analysis. We have implemented the proposed approach for semi-automatic visual subgroup mining in the VIKAMINE (Visual, Interactive and Knowledge-intensive Analysis and Mining Environment) system.

The rest of the paper is organized as follows: We discuss the process model for semi-automatic visual subgroup mining, introduce the general task of subgroup mining and summarize the automatic methods of VIKAMINE in Section 2. Then, we describe the visual subgroup mining approach in Section 3 and discuss related work. In Section 4 we summarize three case studies using data from fielded medical systems in the domain of sonography and in the domain of dental medicine, to demonstrate the applicability and benefit of the approach. We conclude the paper in Section 5 with a discussion of the presented work and we show promising directions for future work.

2 Process Model for Semi-Automatic Visual Subgroup Mining

In the following, we introduce the process model for semi-automatic visual subgroup mining. After that we define the general subgroup mining task, discuss helpful background knowledge, and outline automatic methods supported by VIKAMINE .

The process model is depicted in Figure 1: the dotted boxes represent steps supported by helpful visualization methods as described in Section 3. We distinguish two main tasks: (1) discovering (SD) a set of subgroups, and (2) evaluating and refining the set of the most interesting subgroups. Furthermore, an important initial step, i.e., selecting the relevant objects for analysis – the target and the set of independent variables – can be supported by overview visualizations that show the distributions of variables. In our approach, (incremental) subsequent tuning of these variables is also supported, e.g., utilizing background knowledge as discussed below.

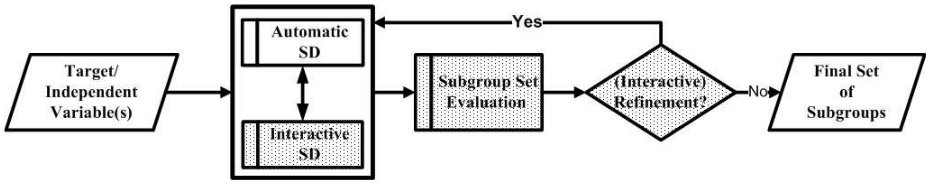

Figure 1: The Semi-Automatic Visual Subgroup Mining Process

In this process both interactive and automatic elements are combined in an active mining approach (c.f., [Motoda, 2002]): the user is directly integrated into the subgroup mining process and can manipulate the subgroup descriptions interactively.

Basic Subgroup Mining

Before defining the subgroup mining task, we first introduce the necessary notions concerning our knowledge representation schema. Let Ω_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.

For a subgroup mining task there are mainly four main properties: the target variable, the subgroup description language, the quality function, and the search strategy. An efficient search method is necessary due to the exponential search space: commonly, beam search is used because of its efficiency [Klösgen, 2002]. The target variable may be binary, nominal or numeric. We will focus on binary target variables – to discover subgroups with a high share 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 included attributes, CB is the case base representing the general population consisting of cases (also called instances), and C specifies (optional) constraints for the mining process. Ω_{SP} denotes the set of all possible subgroup mining problems.

Definition 2 Subgroup Description. A subgroup description $sd = \{e_i\}$ specifies the individuals belonging to the subgroup: it consists of a set of selection expressions (selectors) $e_i = (a_i, V_i)$ that are selections on domains of attributes, where $a_i \in \mathcal{V}_A, V_i \subseteq dom(a_i)$. A subgroup description is defined as the conjunction of its contained selection expressions. We define Ω_{sd} as the set of all possible subgroup descriptions.

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 measures the interestingness of the subgroup and is used by the search method to rank the discovered subgroups during search.

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

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N - n}} \quad (1) \quad , \quad q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)}, \; n \geq \mathcal{T}_{Supp} \quad (2)$$

where p is the relative frequency of the target variable in the subgroup (i.e, the target share of 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) (e.g., [Klösgen, 2002]), 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 minimum coverage threshold $\mathcal{T}_{Supp}$ is necessary.

Background Knowledge for Subgroup Mining

While visualization methods make use of the implicit background knowledge of the user, subgroup mining methods can also utilize explicitly formalized knowledge, e.g., constraints, ontological knowledge, and abstraction knowledge. In the following, we summarize these important knowledge elements and refer to [Atzmueller *et al.*, 2005b] for a 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 attribute *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.

Abnormality/normality information is part of the ontological knowledge that includes information about the domain ontology. 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. 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. Another example for ontological knowledge is given by attribute weights that specify the relative importance of an attribute.

Derived attributes (abstraction knowledge) play a special role in the mining process. These attributes are constructed according to the needs of the user – as rule-based abstractions that can be refined incrementally during the process; they are inferred from basic attributes or other derived attributes. For example, a derived attribute can combine similar attributes to reduce the search space, and to increase the interpretability.

Automatic Subgroup Mining using VIKAMINE

The subgroup mining tool VIKAMINE also includes automatic mining methods besides the visualization methods described in Section 3: several standard search algorithms (e.g. [Klösgen, 2002]) are supported, e.g., beam-search (optionally using strong heuristics for pruning), exhaustive search for constrained search spaces, and *patient* search strategies, e.g., the PRIM algorithm [Friedman and Fisher, 1999].

All these algorithms can utilize background knowledge as introduced above. Additionally, for interactive use of the automatic algorithms, initial subgroups (hypotheses) can be provided for further automatic refinement. Furthermore, subgroups can be clustered to identify similar subgroups (with the same extension, i.e., the same set of covered instances) to find discriminative sets of subgroups. More details can be found in [Atzmueller *et al.*, 2005b].

3 Visual Subgroup Mining

The *Visual Information Seeking Mantra* by [Shneiderman, 1996], "Overview first, zoom and filter, then details-on-demand" is an important guideline for visualization. In an iterative process, the user is able to subsequently concentrate in the interesting data elements by filtering uninteresting data, and focusing (zooming in) on the interesting elements, until finally details are available for an interesting subset of the analyzed elements. We implement this principle in a component (the *zoomtable*) for visualizing the individual distributions of variables: first, the zoomtable can be used to obtain a quick overview of all the variables for analysis. Additionally, it supports the visual search process directly by providing helpful information using visual markers.

Furthermore, we present several techniques for evaluating sets of subgroups. These are visualized for an easier comparison, such that the user is enabled to select the relevant patterns directly in order to identify a small number of highly discriminative and distinct subgroups with a high overall quality. The proposed approach provides a novel integration and effective combination of the individual visualization techniques such that the user can utilize each visualization as needed during the mining process. We illustrate the presented visualization methods by examples from the three case studies presented in Section 4.

3.1 Visualizations for Explorative Subgroup Mining

In this section we present two visualizations for interactive and explorative subgroup mining. First, we discuss the zoomtable that enables an overview of the important variables but can also be used in the subgroup discovery step directly. After that, we describe the *subgroup tuning table* that can be used to optimize a given subgroup by small variations that improve the subgroup – according to objective or subjective quality criteria, e.g., complexity, unusualness or interpretability. In general, the zoomtable shows the value distributions of selected analysis variables in the rows of the table corresponding to the attributes in the first column. Additionally, the zoomtable can contain visual markers for guiding the discovery process that are configurable on the fly.

As an example for a simple configuration, the main component of the zoomtable is depicted in Figure 2. In this figure, the frequencies of the individual values are given below the value names, and the widths of the cells are evenly scaled. Then, a first overview of the value distributions can be obtained. Additionally, the zoomtable can be used for simple correlation analysis similar to basic *OLAP* (Online Analytical Processing) [Han and Kamber, 2000] techniques: in the *sorted* mode, the different rows of the zoomtable can be analyzed w.r.t. other rows, i.e., given a sorting attribute the values of the other attributes are grouped by the values of the respective attribute. An example is shown in Figure 3. In this zoomtable screenshot the cells are not evenly scaled but the width of a cell depicting an attribute value relates to the frequency of the respective attribute value; the values of the attribute "Attachmentloss" are sorted according to the attribute

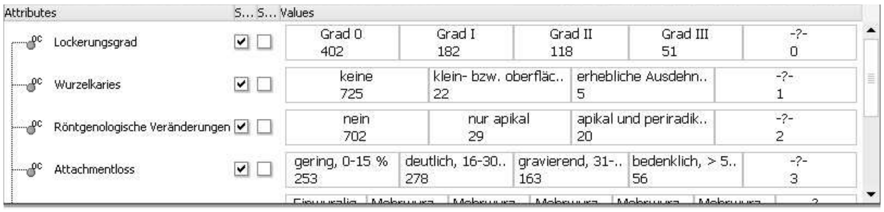

Figure 2: Showing value distributions in the Zoomtable

"Lockerungsgrad" (tooth lax) in the first row. This mechanism can be extended using further sorting attributes such that the values of an attribute are grouped by the several attributes according to a sorting order. Then, the value cells are split up recursively, depending on the attributes that are higher up in the sorting order.

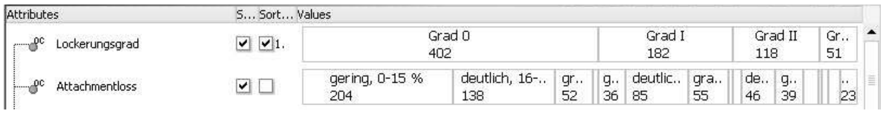

Figure 3: Simple Correlation Analysis in the Zoomtable

For subgroup mining, we utilize additional capabilities provided by the zoomtable. Given a subgroup mining problem defined by the target variable, the set of relevant attributes and additional constraints, the attributes can directly be visualized in the rows of the zoomtable. Then, we can utilize additional information displayed in its cells.

Applying the zoomtable for subgroup mining is exemplified in Figure 4. This screenshot contains the *current subgroup view* (Annotation I) showing the binary target variable "Error Analysis: (EX;;;red; [FP,FN])" that is true for cases where a dental consultation system wrongly inferred a tooth to be extracted; the single selector of the current subgroup is given by "Attachmentloss = gravierend, 31-50%" (attachmentloss = strong). The bars (Annotation II) depict the target distributions in the whole population (upper bar: 108 positives vs. 670 negatives), and in the subgroup (57 vs. 115). The left part of a bar shows the positives, the right part the negative instances. Usually the positives are shown in green color, and the negative instances in red color (in Figure 4 the colors are interchanged due to the error analysis task). 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 limited to the cases covered by the current subgroup; the width of each cell relates to the frequency of the respective attribute value.

For a detailed view, Figure 5 shows the abstract structure of a row of the zoomtable including the type of the attribute, its current ranking, the attribute name, and its value distribution annotated with several visual markers. In general, two of the most impor-

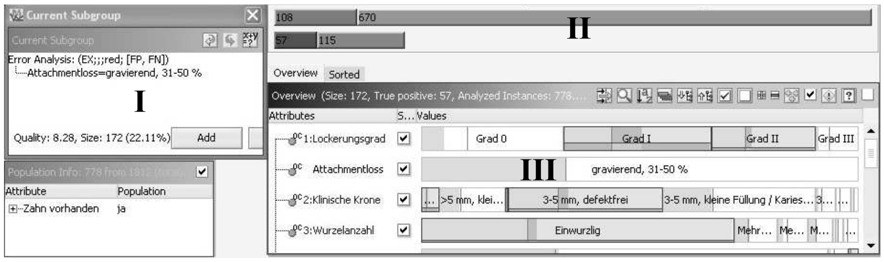

Figure 4: Visual Subgroup Mining using the Zoomtable

tant parameters of a subgroup are the *target share* (p) (c.f., Section 2) and the *size* (n) of the respective subgroup. There is always a trade-off between these parameters that is usually formalized by the applied quality function. So, for the interactive part of the semi-automatic process for subgroup mining, we want to visualize possible future changes or improvements regarding these parameters. The *subgroup size* w.r.t. a future subgroup is given by the width of a specific selector cell. The current target share is visualized in the individual cells by visual markers: (a) indicates the positive and (b)

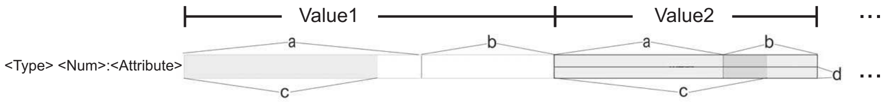

Figure 5: The Zoomtable – Detail View

the negative instances of the current subgroup SG_c; (c) shows the positive instances for the subgroup SG_n, i.e., the subgroup that is constructed by including the particular attribute value. If (c) is larger than (a), then the target share increases adding this selector. Then, (d) shows the relative gain in the target share p, comparing the subgroups SG_c and SG_n, i.e., $d \sim \frac{c-a}{b}$, $c \geq a$, for an easier assessment of small cells. If the height of (d) is zero, then the target share does not increase. If it fills the entire bar, then the target share reaches 100%. We can relate this abstract structure to the table shown in Figure 4, e.g., to the first row showing the attribute "Lockerungsgrad" (tooth lax). Then the annotations in the cell representing the value "Grad I" (minor degree of tooth lax) indicate that the target share significantly increases by adding this selector. So, by interpreting these visual markers which are shown using different colors the user can immediately identify promising improvements of the currently active subgroup.

Furthermore – if enabled – the zoomtable ranks the rows of the table with the most significant improvement, shown by the number in the column left to the value cells in Figure 4 ("Num" in Figure 5). For example, in Figure 6 the best selector "Lockerungsgrad = Grad I" (tooth lax = minor) has been added to the current subgroup, compared to Figure 4. Now, the selector "Klinische Krone = 3-5mm, defektfrei" (crown length = 3-5mm, undamaged) is a potential candidate for further subgroup refinement.

Using such zooming operations, i.e., selecting selectors, the user can manipulate the current subgroup by one click selecting cells in the zoomtable; the zoomtable is animated and updated immediately w.r.t. the respective value distributions. Then, interactive exploration can be performed very easily and effectively.

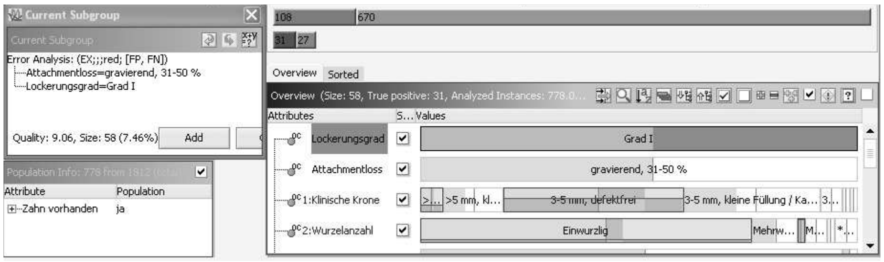

Figure 6: Visual Subgroup Mining using the Zoomtable - "Zooming In" Operation

Figure 7 shows a detailed subgroup tuning table. It is used, if the important (interesting) attribute values have been determined, e.g., using the zoomtable. The user can try out a (user-selected) limited number of choices (de-)selecting attribute values by a single click in a value cell – subgroup specialization and generalization – and immediately assess the impact of the modifications. Slight variations of a subgroup description can be more interesting depending on the preferences and requirements of the user [Atzmueller *et al.*, 2005b]. The individual subgroups are shown in the rows of the table.

Target Variable: Gallstones																					
#	Age			Sex		Liver size						Aorta sclerosis									
	1	2	3	m	f	1	2	3	4	5	6	n	c	Size	TP	FP	Pop.	p0	p	RG	Bin. QF
1			X		X				X	X	X		X	89	37	52	3171	0.172	0.416	1.71	6.17
2			X		X				X	X	X	X	X	119	46	73	3171	0.172	0.387	1.5	6.31
3		X	X		X				X	X	X		X	132	51	81	3171	0.172	0.386	1.5	6.66
4					X			X	X	X	X		X	190	68	122	3177	0.172	0.358	1.3	6.99
5			X		X				X	X	X			207	72	135	3171	0.172	0.348	1.23	6.92
6		X	X		X					X	X	X	X	64	22	42	3171	0.172	0.344	1.2	3.67

Age: 1 = <50, 2 = 50-69, 3 = >=70
Sex: m = male, f = female
Liver size: 1 = smaller than normal,
2 = normal,
3 = marginally increased,
4 = slightly increased,
5 = moderately increased,
6 = highly increased
Aorta sclerosis: n = not calcified, c = calcified

Figure 7: Exemplary Subgroup Tuning Table

Subgroup parameters given in the columns are: (Subgroup) Size, TP (true positives), FP (false positives), Pop. (defined population size), RG (relative gain) and the value of the binomial quality function q_{BT} (Bin. QF), c.f., Section 2. For example, the first line depicts the subgroup (89 cases) described by *Age* $\geq$ 70 *AND Sex=female AND Liver size=*$\{$*slightly or moderately or highly increased*$\}$ *and Aorta sclerosis=calcified* with a target share (gallstones) of 41.6% (p) compared to 17.2% (p_0) in the general population, with a relative gain of 171% (RG).

3.2 Visualizations for Subgroup Comparison and Evaluation

Subgroup mining typically aims to identify a small set of distinct high quality subgroups. Automatic approaches for subgroup selection, i.e., methods that aim to optimize the set of subgroups suffer from several limitations. A subgroup can potentially be described by several different subgroup descriptions, i.e., sets of selectors, if there are multi-correlations between the individual attributes. Then, a representative subgroup needs to be selected from the competing descriptions which is difficult using automatic methods without including background knowledge. Furthermore, subjective quality criteria of the user also need to be taken into account: e.g., the individual trade-off between subgroup size and target, the complexity of a subgroup, the unusualness, and finally the (subjective) interestingness of the subgroup: these criteria are hard to assess without user integration. In the following we describe two types of visualizations for the proposed semi-automatic approach: visualizations for comparing the quality characteristics of subgroups and for assessing hierarchical/redundancy relations between subgroups.

Comparing Subgroup Quality Parameters

For comparing individual subgroups we propose a specialized visualization – the *stacked bar* visualization similar to a spineplot [Theus, 2002]. In this visualization we aim to (1) show the most important subgroup parameters (size n and target share p) that should be (2) easily comparable, in a (3) compact way, such that it is possible to include a high number of subgroups in this visualization.

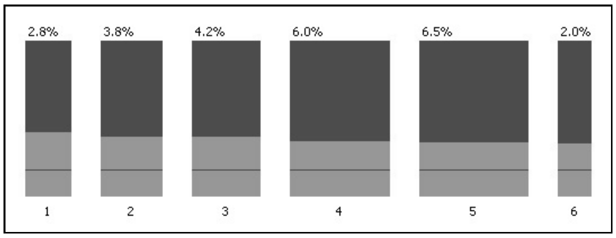

Figure 8: Stacked Bar Visualization

An example of the stacked bar visualization that corresponds to the subgroups shown in Figure 7 (referenced by the numbers in the first column) is shown in Figure 8: the combined area of a stacked bar shows the subgroup size (the relative size w.r.t. the total population is also printed above the bars), and the shares of the positive/negative instances are depicted by the areas of the lower/upper bars. The default target share is indicated by a vertical line within the lower bars which can be compared to the height of the lower bars representing the target share of the subgroup. The stacked bar visualization has the important property that the relevant information for a subgroup, the size and the target share is directly shown in the visualization. Since the bars are arranged horizontally, this visualization is suitable for comparing large sets of subgroups.

Comparing Subgroup Relations

To compare the characteristics of several subgroups, we describe an overview visualization to identify the specialization/generalization relations between subgroups. Furthermore, we provide an overlap/clustering visualization to identify overlapping subgroups, since subgroup mining methods are not necessarily covering approach.

If all selectors of the description of the subgroup A are contained in the description of the subgroup B, then A is a generalization of B and B is a specialization of A. This relation can be visualized as a graph (e.g., [Klösgen and Lauer, 2002]), where the subgroups are the nodes and an edge from A to B exists if B is a specialization of A. The overview visualization is shown in Figure 9. Each bar depicts a subgroup: the left

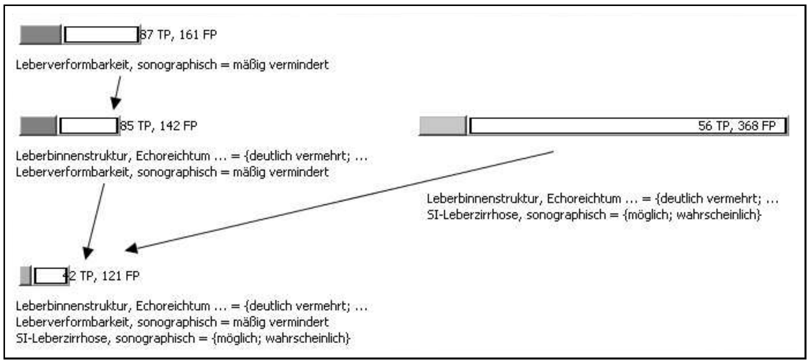

Figure 9: Overview on Sets of Subgroups

sub-bar shows the positives and the right one the negative instances of the subgroup; it is easy to see that the size n is the sum of these parameters, and that the target share is obtained by the fraction of positives and size. The quality of the subgroup is indicated by the brightness of the positive bar such that a darker bar indicates a better subgroup. So, we are able to include the most important parameters, i.e., the size, target share, and the subgroup quality. Since the edges show which subgroups are specializations of other subgroups, it is easily possible to see the effect of additional selectors.

The cluster visualization (given in Figure 10 containing the subgroups of Figure 9) shows the overlap of subgroups (i.e., their similarity) and can be used to detect redundant subgroups. For example, if all positive instances of a subgroup A are also contained in another subgroup B with less negative instances, then the subgroup A is potentially redundant. The similarity $SGSim(s_1, s_2)$ of two subgroups s_1 and s_2 can be defined using the intersection and the union of their instances:

$$SGSim(s_1, s_2) = \frac{\mid s_1 \cap s_2 \mid}{\mid s_1 \cup s_2 \mid}.$$

Thus, the cluster visualization can also show redundancy between subgroups that are not similar w.r.t. their descriptions but only similar concerning the covered instances.

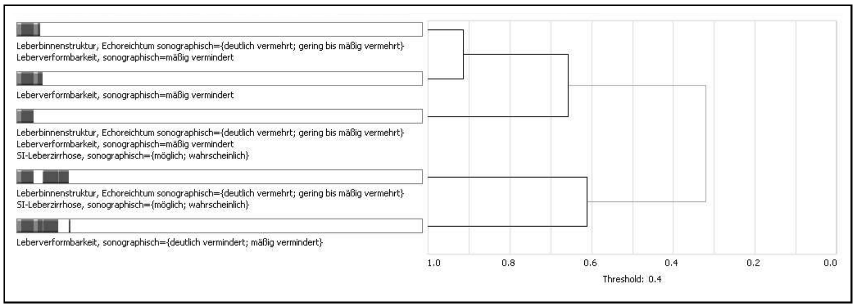

Figure 10: Visualizing Overlap/Clusters of Subgroups

To indicate overlapping subgroups the cases are arranged in the same order for each row corresponding to a subgroup. If an instance is contained in a subgroup, it is marked in green, if it is positive, and red if it is negative, w.r.t. the target variable. To determine the individual clusters, we apply a bottom-up hierarchical complete-linkage clustering algorithm (e.g., [Han and Kamber, 2000]), starting with the single subgroups and merging the two most similar clusters recursively, as shown in Figure 10. The process terminates if a certain similarity threshold is reached. This threshold can also be determined automatically using an adapted algorithm from [Kirsten and Wrobel, 1998].

3.3 Related Work

There are several approaches that support the visual analysis of dependencies between attributes: *Attribute Explorer* [Spence and Tweedie, 1998] uses a bar-chart visualization of the attributes and visualizes where certain constraints of the attributes fail. The *InfoZoom* system [Spenke, 2001] visualizes the value distributions of attributes in single rows of a table, and allows *zooming in* on individual values. Our approach was inspired by this idea but extends it significantly, since we also guide the user during the subgroup mining process by visualizing additional quality parameters such as the future target share, and the gain of a specific selector directly in the zoomtable. Changes in the zoomtable are also visualized with respect to the current subgroup dynamically.

For comparing subgroups, several visualizations have already been proposed (c.f., [Gamberger *et al.*, 2002]), e.g., displaying subgroups using circles or boxes. In contrast to these, the presented stacked bar visualization includes all the relevant parameters and enables their comprehensive comparison which is often problematic for the other approaches, e.g., for small subgroup sizes.

The specialization hierarchy is inspired by a similar graph by [Klösgen and Lauer, 2002] but there the quality of a subgroup is not visualized directly. In addition, we also include other relevant parameters, e.g., the target share in this visualization. By visualizing the positive and negative instances of each subgroup in the graph it is possible to see the direct effect of additional selectors in the respective subgroup descriptions.

4 Case Studies

The first two case studies were performed in the medical domain of sonography: we used subgroup mining for knowledge discovery, i.e. discovering subgroups in which the presence of a disease is significantly more probable than in the whole population. The second case study aimed to identify characteristic profiles of examiners for quality control. The third case study concerns the domain of dental medicine were we used subgroup mining for interactive knowledge refinement of a knowledge-based system.

4.1 Case Studies – Knowledge Discovery and Quality Control

For the first two case studies we used cases taken from the SONOCONSULT system [Huettig *et al.*, 2004] – a medical documentation and consultation system for sonography – that is in routine use. The applied SONOCONSULT case base for the first case study contains 4358 cases, for the second case study we used 7096 cases. The domain ontology contains about 560 attributes with about 5 symbolic values on average. This indicates the potentially huge search space for subgroup discovery.

For both case studies, we first applied automatic subgroup mining methods. It turned out, that too many results were discovered; most of these were not regarded as (clinically) interesting or were already known to the domain specialist. Therefore, the domain specialist directly used the VIKAMINE system, both the visual and automatic components in a semi-automatic approach. Automatic methods were used to refine hypotheses, and to provide initial starting points for further analysis. Additionally, background knowledge could be integrated very effectively to focus the search on the subset of interesting patterns. The discovered subgroups were evaluated by domain specialists according to (clinical) novelty, interestingness, and actionability aspects.

The first case study applied subgroup mining for knowledge discovery, i.e., discovering characteristic subgroups for certain diseases, e.g., the subgroup *Age* $\geq$ 70 *AND Sex=female AND Liver size=*$\{$*slightly or moderately or highly increased*$\}$ *and Aorta sclerosis=calcified* for the target variable *Gallstones*. Figures 7 and 8 show some further (clinically) interesting subgroups that were discovered.

The second case study applied subgroup mining for quality control concerning documentation purposes. Since sonographic examination and documentation is highly dependent on the skills of the examiners, there are significant deviations concerning their documentation behavior. If these deviations can be identified they can potentially be utilized for training purposes. To identify such deviations the subgroup mining method is used to discover documentation patterns, i.e., certain symptom combinations that are observed significantly more (in-)frequently in conjunction with certain examiners. Examples of the discovered profiles are shown in Figures 9 and 10, e.g., the subgroup *Liver Plasticity=moderately reduced AND Liver Echogenicity = moderately or strongly increased*. For a more detailed description we refer to [Atzmueller *et al.*, 2005b] and [Atzmueller *et al.*, 2005c], respectively.

Applying the process model, the domain specialists considered the visualization component very helpful, since it enabled an easy step by step analysis: single factors could be identified first, and then the respective subgroups were refined. This rapidly enabled first positive results for the domain specialists. Furthermore, subgroups discovered by the automatic search method were also validated and refined interactively. The domain specialist considered it extremely helpful that the process provided effective incremental feedback improving promising initial results.

4.2 Subgroup Mining for Knowledge Refinement

The third case study was performed in the domain of dental medicine implemented with a consultation and documentation system for dental findings regarding any kind of prosthetic appliance. The system has been developed in cooperation with the department of prosthodontics at the Würzburg University Hospital. In the first level the system proposes the teeth that could be conserved and the teeth that should be extracted, providing the system solution. For knowledge refinement we try to identify subgroups of incorrectly solved cases using a binary target variable that is true, if the system solution differs from the correct solution provided by a domain specialist. Then, we consider the factors that describe the subgroup as indicators for changes to the knowledge base.

The semi-automatic process is essential for the interactive knowledge refinement task: since we also experienced incorrect case descriptions purely automatic refinement methods were not applicable since they require a correct case base. The domain specialist applying the VIKAMINE system was able to use the automatic methods to obtain first coarse hypotheses. These were then checked and refined using the interactive tools. Examples for the application of the subgroup mining process are given in Figure 4 and Figure 6. The method was very well accepted by the domain specialist, who was able to directly inspect and change the subgroups and cases by himself. Applying proposed changes to the knowledge base turned out quite successful since we were able to reduce the error rate by 50%. We refer to [Atzmueller *et al.*, 2005a] for further details.

5 Conclusion and Outlook

In this paper, we presented a novel semi-automatic approach for visual subgroup mining implemented in the VIKAMINE system. We discussed the general subgroup mining process, and described how it can be improved by user integration utilizing appropriate visualization methods. We discussed the zoomtable as the proposed key visualization technique to guide the subgroup mining process, and additionally presented several visualizations for subgroup comparison and evaluation. We shortly discussed three case studies that showed the applicability and benefit of the presented approach.

In the future, we plan to consider a combination of automatic and interactive techniques for subgroup post-processing, e.g., combining automatic methods for causal analysis of subgroups and appropriate visualizations for this task.

References

[Atzmueller *et al.*, 2005a] Martin Atzmueller, Joachim Baumeister, Achim Hemsing, Ernst-Jürgen Richter, and Frank Puppe. Subgroup Mining for Interactive Knowledge Refinement. In *Proc. 10th Conference on Artificial Intelligence in Medicine (AIME 05)*, LNAI 3581, pages 453–462, Berlin, 2005. Springer.

[Atzmueller *et al.*, 2005b] 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)*, pages 647–652, 2005.

[Atzmueller *et al.*, 2005c] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. Profiling Examiners using Intelligent Subgroup Mining. In *Proc. 10th Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005)*, pages 46–51, 2005.

[Fayyad *et al.*, 1996] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padraic Smyth. From Data Mining to Knowledge Discovery: An Overview. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, 1996.

[Friedman and Fisher, 1999] J. Friedman and N. Fisher. Bump Hunting in High-Dimensional Data. *Statistics and Computing*, 9(2), 1999.

[Gamberger *et al.*, 2002] Dragan Gamberger, Nada Lavrac, and Dietrich Wettschereck. Subgroup visualization: A method and application in population screening. In *Proc. 7th Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2002)*, 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.

[Han and Kamber, 2000] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher, 2000.

[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.

[Kirsten and Wrobel, 1998] M. Kirsten and S. Wrobel. Relational Distance-Based Clustering. In D. Page, editor, *Proc. Conference ILP 98*, volume 1446 of *LNAI*, pages 261 – 270, 1998.

[Klösgen and Lauer, 2002] Willi Klösgen and Stephan R. W. Lauer. *Handbook of Data Mining and Knowledge Discovery*, chapter 20.1: Visualization of Data Mining Results. Oxford University Press, New York, 2002.

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

[Motoda, 2002] Hiroshi Motoda. Active Mining, A Spiral Model of Knowledge Discovery (Invited talk). In *Proc. 2002 IEEE Intl. Conference on Data Mining*, Maebashi City, Japan, 2002.

[Shneiderman, 1996] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE Computer Society, 1996.

[Spence and Tweedie, 1998] Robert Spence and Lisa Tweedie. The Attribute Explorer: Information Synthesis via Exploration. *Interacting with Computers*, 11(2):137–146, 1998.

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

[Theus, 2002] Martin Theus. *Handbook of Data Mining and Knowledge Discovery*, chapter 15.2: Highly Multivariate Interaction Techniques. Oxford University Press, New York, 2002.

[Wills and Keim, 2002] Graham J. Wills and Daniel Keim. *Handbook of Data Mining and Knowledge Discovery*, chapter 15.1: Interactive Statistical Graphics. Oxford University Press, New York, 2002.

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