

# Semi-Automatic Refinement and Assessment of Subgroup Patterns

Martin Atzmueller and Frank Puppe

University of Würzburg  
Department of Computer Science VI  
Am Hubland, 97074 Würzburg, Germany  
{atzmueller, puppe}@informatik.uni-wuerzburg.de

## Abstract

This paper presents a methodological approach for the semi-automatic refinement and assessment of subgroup patterns using summarization and clustering techniques in the context of intelligent data mining systems. The method provides the suppression of irrelevant (and redundant) patterns and facilitates the intelligent refinement of groups of similar patterns. Furthermore, the presented approach features intuitive visual representations of the relations between the patterns, and appropriate techniques for their inspection and evaluation.

## Introduction

Intelligent data mining systems are commonly applied to obtain a set of *novel, potentially useful*, and ultimately *interesting* patterns from a given (large) data set (Fayyad, Piatetsky-Shapiro, & Smyth 1996). However, one of the major problems for standard data mining techniques is given by a large set of potentially interesting patterns that the user needs to assess. In addition, especially the application of descriptive data mining techniques like methods for association rule mining (Agrawal & Srikant 1994) or subgroup mining, e.g., (Wrobel 1997; Klösgen 1996; Atzmueller, Puppe, & Buscher 2005), often yields a very large set of (interesting) patterns. Then, the set of the discovered patterns needs to be evaluated and validated by the user in order to obtain a set of relevant patterns. In such scenarios, a naive (interactive) browsing approach often cannot cope with such a large number of patterns.

Consider a busy end-user that needs to assess a large set of potentially interesting patterns. In order to identify a set of relevant patterns all the proposed potentially interesting patterns need to be browsed, evaluated, compared, and validated. A given pattern might just be a specialization of an interesting pattern, both having the same interestingness value as rated by a quality function. Furthermore, an interesting (statistical) phenomenon might syntactically be described by two competing descriptions.

In such cases, the set of patterns is redundant and the browsing effort needs to be facilitated using filtering techniques and intelligent summarization methods: (Syntactically) irrelevant patterns should be suppressed, while similar patterns should be grouped or clustered appropriately.

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

There exist knowledge-intensive methods, e.g., (Atzmueller, Puppe, & Buscher 2005), for focusing the applied discovery method on the set of relevant patterns. Such approaches, for example, can reduce the search space, focus the search method, and increase the representational expressiveness of the discovered set of patterns significantly. However, the set of the discovered patterns can still be quite large for huge search spaces.

Essentially, the *assessment* of the patterns, i.e., the evaluation and validation of the patterns in order to determine their final usefulness and interestingness need to be facilitated by specialized techniques, c.f., (Fayyad, Piatetsky-Shapiro, & Smyth 1996; Gomez-Perez, Juristo, & Pazos 1995). In this paper, we propose visual methods applying clustering techniques in order to refine a set of mined patterns: The set of patterns needs to be summarized in order to be more concise for the user, and also to provide a better overview on the (reduced) set of patterns. We propose two techniques for that task: 1) An intelligent filtering approach and 2) special clustering methods.

Since the interestingness of the patterns also significantly depends on subjective interestingness measures, usually semi-automatic approaches are preferred compared to purely automatic methods. The *Visual Information Seeking Mantra* by (Shneiderman 1996), 'Overview first, zoom and filter, then details-on-demand' is an important guideline for visualization methods. In an iterative process, the user is able to subsequently concentrate on the interesting data by filtering irrelevant and redundant data, and focusing (zooming in) on the interesting elements, until finally details are available for an interesting subset of the analyzed patterns. The described techniques are integrated into an incremental process model featuring the following steps:

- A pattern filtering step for suppressing redundant patterns.
- A clustering method for generating a comprehensive view on a refined set of (diverse) patterns.
- Specialized visual techniques for the inspection of the refined patterns, that can be incrementally applied.

The proposed process covers the whole range after pattern discovery towards pattern evaluation and validation. The visualization techniques can be applied during the intermediate steps in order to inspect the intermediate results in detail. Furthermore, they can be applied for analyzing the refined set of patterns during the evaluation and validation phase.

The rest of the paper is organized as follows: We briefly introduce the background of descriptive data mining exemplified by subgroup mining and association rule mining. After that, we present the process model for the semi-automatic refinement and assessment approach. We describe the integrated methods in detail and illustrate them using examples from the medical domain of sonography. Finally, the results of the paper are discussed and concluded.

## Background: Subgroup Patterns

In the context of this work, we focus on subgroup mining methods, e.g., (Wrobel 1997; Klösgen 1996; Atzmueller, Puppe, & Buscher 2005), for discovering interesting patterns. In the following section we introduce subgroup patterns, and show their relation to the well-known representation of association rules, e.g., (Agrawal & Srikant 1994).

Subgroup patterns, often provided by conjunctive rules, describe 'interesting' subgroups of cases/instances, e.g., "the subgroup of 16-25 year old men that own a sports car are more likely to pay high insurance rates than the people in the reference population." The main application areas of subgroup mining are exploration and descriptive induction, to obtain an overview of the relations between a target variable and a set of explaining variables. The exemplary subgroup above is then described by the relation between the independent (explaining) variables (Sex = male, Age  $\leq$  25, Car = sports car) and the dependent (binary) target variable (Insurance Rate = high). The independent variables are modeled by selection expressions on sets of attribute values.

Let  $\Omega_A$  denote the set of all attributes. For each attribute  $a \in \Omega_A$  a range  $dom(a)$  of values is defined. An attribute-value assignment  $a = v$ , where  $a \in \Omega_A, v \in dom(a)$ , is called a *feature*. We define the feature space  $\mathcal{V}_A$  to be the (universal) set of all features. A single-relational propositional *subgroup description* is defined as a conjunction

$$sd = e_1 \wedge e_2 \wedge \dots \wedge e_n$$

of (extended) features  $e_i \subseteq \mathcal{V}_A$ , which are then called selection expressions, where each  $e_i$  selects a subset of the range  $dom(a)$  of an attribute  $a \in \Omega_A$ . We define  $\Omega_{sd}$  as the set of all possible subgroup descriptions. The subgroup size  $n(s)$  for a subgroup  $s$  is determined by the number of instances/cases covered by the subgroup description  $sd$ . For a binary target variable, we define the true positives  $tp(sd)$  as those instances containing the target variables and the false positives  $fp(sd)$  as those instances not containing the target variable, respectively.

In contrast to subgroup patterns, an *association rule*, e.g., (Agrawal & Srikant 1994), is given by a rule of the form  $sd_B \rightarrow sd_H$ , where  $sd_B$  and  $sd_H$  are subgroup descriptions; the rule body  $sd_B$  and the rule head  $sd_H$  specify sets of items. For an insurance domain, for example, we can consider an association rule showing a combination of potential risk factors for high insurance rates and accidents:

$$\begin{aligned} & \text{Sex} = \text{male} \wedge \text{Age} \leq 25 \wedge \text{Car} = \text{sports car} \\ & \rightarrow \text{Insurance Rate} = \text{high} \wedge \text{Accident Rate} = \text{high} \end{aligned}$$

A *subgroup pattern* is thus a special association rule, namely a horn clause  $sd \rightarrow e$ , where  $sd \in \Omega_{sd}$  is a subgroup description and the feature  $e \in \mathcal{V}_A$  is called the target variable.

Considering association rules, the quality of a rule is commonly measured by its support and confidence, and the data mining process searches for association rules with arbitrary rule heads and bodies. For subgroup patterns there exist various (more refined) quality measures, e.g., (Klösgen 1996; Atzmueller, Puppe, & Buscher 2005): Since an arbitrary quality function can be applied, the anti-monotony property of support used in association rule mining cannot be utilized in the general case. E.g., the used quality function can combine the difference of the confidence (target share) and the apriori probability of the target variable with the size of the subgroup (given by the number of covered instances). Since mining for interesting subgroup patterns is more complicated, usually a fixed binary target variable is provided as input to the discovery process.

## The Semi-Automatic Refinement Process

Since the mining process usually results in a very large set of patterns, appropriate refinement techniques need to be applied. Commonly, quality functions for rating a certain pattern are utilized in order to rate their interestingness. However, even after the application of such quality functions the set of interesting patterns can still be quite large. In addition, such a quality function is only able to cover objective quality criteria. Since there are also additional subjective quality criteria, these need to be captured by the user directly, e.g., using suitable visualization techniques.

The refinement of the mined patterns serves three main purposes: It enables a better presentation of the results due to decreasing the redundancy between the individual patterns, and it also provides for a better overview of the pattern. Then, the user is able to determine more easily what is relevant according to the given analysis goals. The set of patterns can then be comprehensibly assessed applying the the proposed refinement process shown in Figure 1. It consists of the following steps that can be applied in an incremental cycle until the final results are obtained:

1. **Pattern Filtering:** The set of mined patterns is reduced to a set of condensed patterns that convey the same information as the initial set of patterns.
2. **Pattern Clustering:** The obtained patterns are clustered according to their overlap relations in order to determine clusters that are similar with respect to their contained instances/cases. In this way, sets of patterns can be identified that are described by different syntactic descriptions but covering (approximately) the same (or similar) sets of instances. Thus, patterns can be identified that describe different phenomena in different ways.
3. **Evaluation and Validation:** Finally, the patterns need to be evaluated and validated by the user. This usually depends on the analysis goals and the background knowledge of the analyst. Using appropriate visualization techniques, the obtained patterns and clusters can be intuitively analyzed and inspected in detail. Such visualization steps can also always be applied during the intermediate steps.

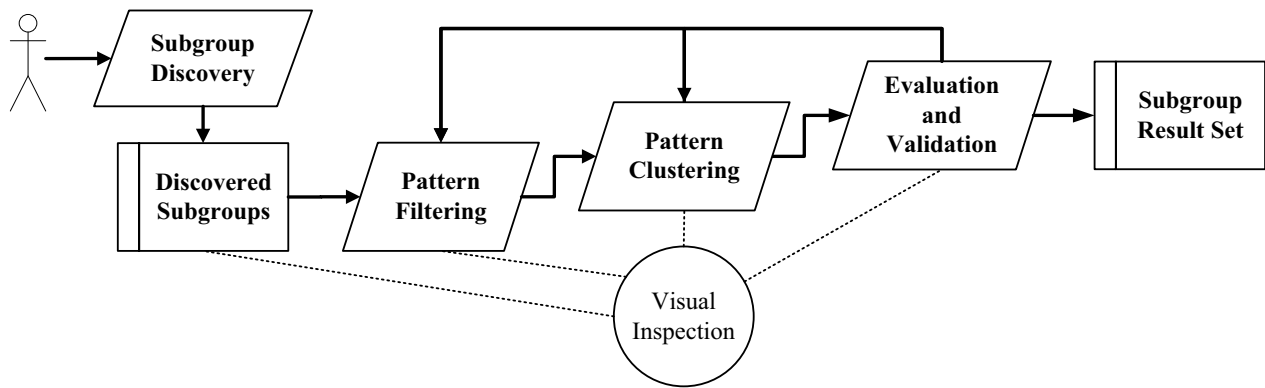


Figure 1: Process model for the semi-automatic refinement and assessment of subgroup patterns.

In the following, we describe the techniques for pattern filtering and clustering in detail, before we show the visualization methods (Atzmueller & Puppe 2005) for assessing the intermediate and/or final results of the process.

### Pattern Filtering

Due to multi-correlations between the independent variables, the discovered subgroups can overlap significantly. Often subgroups can be described by several competing subgroup descriptions, i.e., by disjoint sets of selection expressions. Furthermore, subgroups can also often be described by several overlapping descriptions, i.e., for which the selection expressions contained in the subgroup description have an inclusion relation to each other.

In the following, we focus on the issue of *condensed representations*, e.g., closed itemsets, c.f., (Pasquier *et al.* 1999). For example, closed frequent-set based approaches allow the reconstruction of all frequent sets given only the closed sets. In general, closure systems aim at condensing the given space of patterns into a reduced system of relevant patterns that formally convey the same information as the complete space. Intuitively, closed itemsets can be seen as maximal sets of items covering a maximal set of examples. Based upon the definition of closed-itemsets, e.g., (Pasquier *et al.* 1999), we can define closed subgroup descriptions using the subgroup size  $n$  as follows:

**Definition 1 (Closed Subgroup Description)** A subgroup description  $sd \in S$  is called closed with respect to a set  $S$  if there exists no subgroup description  $sd' \in S$ ,  $sd' \supset sd$ , for which  $n(sd) = n(sd')$ .

If only the closed subgroup descriptions are considered, then for equivalent and equally-sized subgroups, the subgroup with the longest subgroup description is selected.

(Garriga, Kralj, & Lavrac 2006) have shown that raw closed sets can be adapted for labeled data: For discriminative purposes we can then contrast the covering properties on the positive and the negative cases. Concerning subgroup discovery, we can specialize this definition for the target class cases contained in the subgroup, i.e., the true positives.

**Definition 2 (Target-Closed Subgroup Description)** A subgroup description  $sd \in S$  is called target-closed with respect to a set  $S$  if there exists no subgroup description  $sd' \in S$ ,  $sd' \supset sd$ , for which  $tp(sd) = tp(sd')$ .

For redundancy management, we can focus on the target-closed subgroup descriptions: This can be intuitively explained by the fact, that if  $sd \subseteq sd'$  then  $tp(sd') \subseteq tp(sd)$  and  $fp(sd') \subseteq fp(sd)$ . So, focusing on the target class cases  $tp$ , if  $tp(sd') = tp(sd)$ , then either  $sd'$  covers the same set of negatives, or it is even better concerning its discriminative power since it covers less negatives than  $sd$ .

Additionally, we can further reduce the set of target-closed subgroup descriptions by considering the overlap of the false positives of each description (Garriga, Kralj, & Lavrac 2006): Considering two target-closed subgroup descriptions  $sd \subset sd'$  for which  $fp(sd) = fp(sd')$  we can remove the longer subgroup description  $sd'$  since we can then conclude that  $tp(sd') \subset tp(sd)$ .

Then, we have three options for the pattern filtering step: We can first present the target-closed subgroup descriptions only, since these are the relevant ones concerning the target class cases. Next, we can reduce this set further by removing (irrelevant) subgroups that have the same coverage on the negatives but a reduced coverage on the target-class cases.

Finally, we can add some boundary information to the target-closed subgroup description by including the *minimal generator* of the target-closed subgroup description: A generator is a non-target-closed subgroup description that is equivalent to its target-closed subgroup description with respect to the covered cases. Annotating the closed subgroup descriptions with their minimal generator can then help the user for obtaining a better overview on the structure of the individual closed descriptions: The minimal generator provides a boundary for the shortest equivalent subgroup description corresponding to the closed description. If the user is interested in more compact descriptions, e.g., for discrimination, then usually the shorter subgroup descriptions are more interesting. In contrast, for characterization purposes often the longer descriptions provide more insight.

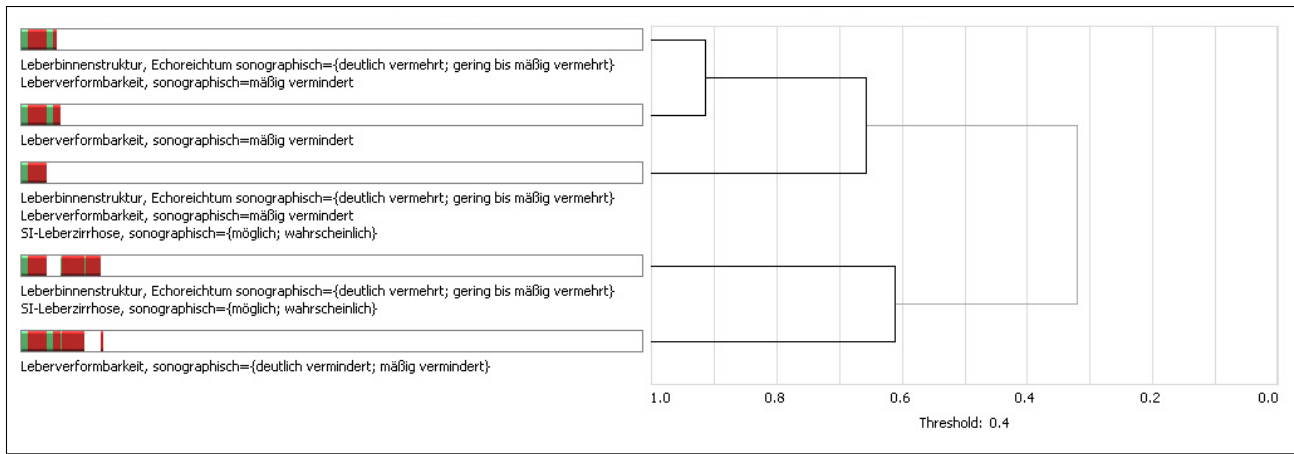


Figure 2: Visualizing overlap/clusters of subgroups. The clustered exemplary subgroups contain the attributes liver plasticity ("Leber Verformbarkeit"), liver echogenicity ("Leber Binnenstruktur, Echoreichtum"), and liver cirrhosis ("SI-Leberzirrhose").

### Pattern Clustering

Clustering subgroups enables their summarization based on their similarity. Individual clusters are defined according to a specified minimal similarity, or they can be automatically constructed using a quality function for the clusters. While grouping and ordering sets of subgroups preserves their redundancy, it enables the user to inspect the (ordered) set of subgroups and to discover hidden relations between the subgroups, i.e., alternative descriptions and multi-correlations between selectors.

Clustering is performed utilizing a similarity measure for pairs of subgroups. A simple symmetric similarity measure is based on the overlap of a pair of subgroups  $s_i, s_j$ , given by the fraction of the intersection and the union size of the covered instances/cases:

$$sim(s_i, s_j) = \frac{|s_i \cap s_j|}{|s_i \cup s_j|}.$$

It is easy to see that  $sim(s_i, s_j) = 0$  for disjoint subgroups and 1.0 for equal subgroups. Then, a bottom-up hierarchical complete-linkage clustering algorithm, e.g., (Han & Kamber 2000, Ch. 8.5), is applied. We start with the single subgroups and merge the two most similar clusters recursively using an algorithm adapted from (Kirsten & Wrobel 1998), as shown in Figure 2. The aggregation process terminates if a certain similarity threshold, i.e., the *split similarity*, is reached, that can be specified by the user.

In order to determine this threshold automatically, cluster-quality functions need to be selected. Since usually large clusters with a high intra-cluster similarity are desired, the quality function should assign a high quality value to a set of clusters with a low inter-cluster similarity and a high intra-cluster similarity. In the presented approach, we apply the following quality function to determine the split similarity  $split$  for a set of clusters  $C(sim)$  corresponding to a certain subgroup similarity  $sim$ :

$$split = \operatorname{argmax}_{sim} \frac{1}{|C(sim)|} \sum_{c \in C(sim)} iSim(c) * \log(|c|).$$

This quality function trades off the intra-cluster similarity  $iSim(c)$  of a cluster  $c$  and the number of subgroups included in the cluster  $c$ .

When the clustering results are analyzed interactively the user can select representative descriptions from the clusters, e.g., a minimal set of the most frequent selectors occurring in all definitions of the subgroups of a cluster. Alternatively, either the cluster *kernel*, i.e., the intersection of all subgroups contained in the cluster, or the cluster *extension*, i.e., the union of all subgroups can be considered. These can also be represented as subgroups (described by a complex selection expression) themselves.

### Pattern Evaluation and Validation

The final decision whether the discovered patterns are novel, interesting and potentially useful has to be made by the domain specialist, e.g., (Ho *et al.* 2002). Therefore, pattern evaluation by the user is essential. We propose interactive techniques for pattern evaluation and analysis that support the user in assessing the subjective quality criteria of a subgroup. The introspection and analysis techniques are orthogonal to the common presentation and integration steps applied in subgroup mining. Similar to the mining results themselves, the characteristics need to be easily understandable and transparent for the user. In the following section we discuss the visualization techniques that are applied for subgroup evaluation and analysis.

Pattern validation is an important step after the data mining process has been performed. Moreover, it is essential in order to verify that the discovered patterns are not only due to spurious associations: If a lot of statistical tests are applied during the data mining process, then this may result in the erroneous discovery of 'significant' subgroups due to the statistical *multiplicity effect*. Then, correction techniques, e.g., a Bonferroni-adjustment (Jensen & Cohen 2000), need to be applied. The patterns can also be validated using an independent test set, or by performing prospective studies, e.g., in the medical domain.

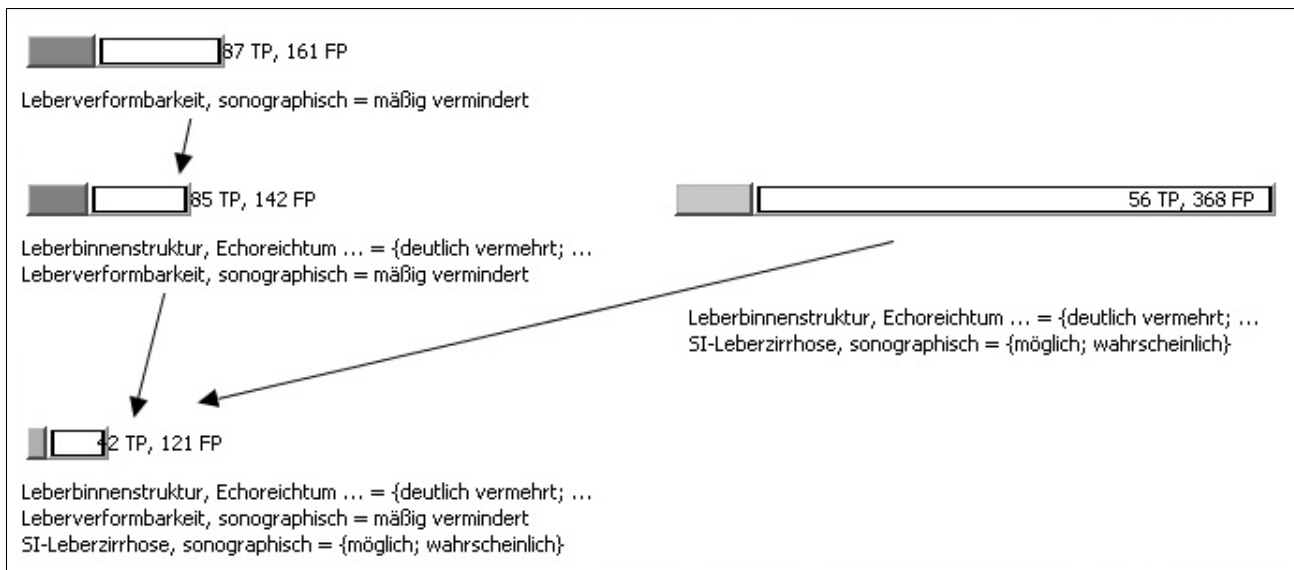


Figure 3: Overview on sets of subgroups: The shown subgroup specializations consider the attributes Liver plasticity ("Leber Verformbarkeit"), liver echogenicity ("Leber Binnenstruktur, Echoreichtum"), and liver cirrhosis ("SI-Leberzirrhose").

### Integrating Visualization Techniques

In the following, we describe three visualizations for subgroup refinement and analysis (Atzmueller & Puppe 2005). According to the *Visual Information Seeking Mantra* by (Shneiderman 1996), 'Overview first, zoom and filter, then details-on-demand', we present an overview visualization and a clustering visualization for obtaining an initial view on the mined patterns. Then, sets of subgroups, e.g., contained in specific clusters can be analyzed in detail.

**Overview visualization** The overview visualization is shown in Figure 3. Each bar depicts a subgroup: The left sub-bar shows the positives and the right one the negative instances of the subgroup; it is easy to see that the subgroup 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 in the visualization, 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. By visualizing the positive and negative instances of each subgroup in the graph it is possible to see the direct effect of these additional selectors in the respective subgroup descriptions. Furthermore, we can add the patterns that are obtained by generalizing the subgroups contained in the overview visualization in all possible ways, in order to see the effect of all possible selector combinations.

**Cluster Visualization** The cluster visualization (depicted in Figure 2 containing the subgroups of Figure 3) shows the overlap of subgroups, i.e., their similarity. It can be used to detect redundant subgroups, e.g., if (approximately) all pos-

itive 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 of two subgroups  $s_1$  and  $s_2$  is defined using a symmetric similarity measure taking the intersection and the union of the respective subgroup instances into account, as defined above

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, with respect to the target variable. Thus, the cluster visualization also shows the redundancy between subgroups that are not similar with respect to their descriptions but only similar concerning the covered instances. If a potentially redundant subgroup is not a specialization of the non-redundant subgroup, i.e., its description is different, then the application of additional semantic criteria might be needed in order to infer if the subgroup is really redundant.

**Subgroup Detail Visualization** Figure 4 shows a detailed visualization of a set of subgroups in the form of a tabular representation. The individual subgroups are shown in the rows of the table: The subgroup description is given by a set of selected columns corresponding to the attribute values. Next, the subgroup parameters are shown, which include the (subgroup) Size, TP (true positives), FP (false positives), Pop. (defined population size), RG (relative gain), and the value of the applied quality function (Bin. QF).

Besides applying this visualization for the inspection of subgroups, it can also be used for subgroup refinement: Given a set of attributes, the user can select each attribute value as a selector for specialization by a single click in a value cell. In this manner, the subgroup description can be fine-tuned by the user, and subgroup specialization and generalization operations can be performed very intuitively.

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	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	X	132	51	81	3171	0.172	0.386	1.5	6.66
4					X		X	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	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 4: Exemplary subgroup detail visualization: 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.

## Conclusions

In this paper we have presented a methodological approach for the semi-automatic refinement and assessment of subgroup patterns. We have described techniques for suppressing irrelevant patterns, for clustering a set of subgroups in order to improve their assessment by the user, and we have shown several visualizations for the interactive analysis and evaluation of a set of subgroup patterns that can be applied throughout the refinement and assessment process. The approach was illustrated by real-world examples from the medical domain of sonography.

In the future we are planning to consider extended visualizations for the refinement of subgroup patterns. Another promising extension of the presented work is given by automatic methods for the selection of relevant subgroups from the clusters: An interesting direction is given by applying quality functions to rate sets of subgroups, c.f., (Atzmueller, Baumeister, & Puppe 2004) for rating sets/combinations of rules.

## Acknowledgements

This work has been partially supported by the German Research Council (DFG) under grant Pu 129/8-1.

## References

- Agrawal, R., and Srikant, R. 1994. Fast Algorithms for Mining Association Rules. In Bocca, J. B.; Jarke, M.; and Zaniolo, C., eds., *Proc. 20th Int. Conf. Very Large Data Bases, (VLDB)*, 487–499. Morgan Kaufmann.
- Atzmueller, M., and Puppe, F. 2005. Semi-Automatic Visual Subgroup Mining using VIKAMINE. *Journal of Universal Computer Science (JUCS), Special Issue on Visual Data Mining* 11(11):1752–1765.
- Atzmueller, M.; Baumeister, J.; and Puppe, F. 2004. Quality Measures for Semi-Automatic Learning of Simple Diagnostic Rule Bases. In *Proc. 15th Intl. Conference on Applications of Declarative Programming and Knowledge Management (INAP 2004)*, 203–213.
- Atzmueller, M.; Puppe, F.; and Buscher, H.-P. 2005. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, 647–652.
- Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining*. AAAI Press. 1–34.
- Garriga, G. C.; Kralj, P.; and Lavrac, N. 2006. Closed Sets for Labeled Data. In *Proc. 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, 163–174. Berlin: Springer Verlag.
- Gomez-Perez, A.; Juristo, N.; and Pazos, J. 1995. Evaluation and Assessment of the Knowledge Sharing Technology. In *Towards Very Large Knowledge Bases*, 289–296. IOS Press.
- Han, J., and Kamber, M. 2000. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher.
- Ho, T.; Saito, A.; Kawasaki, S.; Nguyen, D.; and Nguyen, T. 2002. Failure and Success Experience in Mining Stomach Cancer Data. In *Intl. Workshop Data Mining Lessons Learned, Intl. Conf. Machine Learning*, 40–47.
- Jensen, D. M., and Cohen, P. R. 2000. Multiple Comparisons in Induction Algorithms. *Machine Learning* 38(3):309 – 338.
- Kirsten, M., and Wrobel, S. 1998. Relational Distance-Based Clustering. In Page, D., ed., *Proc. Conference ILP 98*, volume 1446 of *LNAI*, 261 – 270.
- Klösgen, W. 1996. Explora: A Multipattern and Multi-strategy Discovery Assistant. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R., eds., *Advances in Knowledge Discovery and Data Mining*. AAAI Press. 249–271.
- Pasquier, N.; Bastide, Y.; Taouil, R.; and Lakhal, L. 1999. Discovering Frequent Closed Itemsets for Association Rules. In *Proc. 7th Intl. Conference on Database Theory (ICDT 99)*, volume 1540 of *Lecture Notes in Computer Science*, 398–416. Springer.
- Shneiderman, B. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. IEEE Symposium on Visual Languages*, 336–343.
- Wrobel, S. 1997. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, 78–87. Berlin: Springer Verlag.