Towards Knowledge-Intensive Subgroup Discovery

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

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

Abstract

Subgroup discovery can be applied for exploration or descriptive induction in order to discover "interesting" subgroups of the general population, given a certain property of interest. In domains with available background knowledge, the user usually wants to utilize this to improve the quality of the subgroup discovery results.

We describe a knowledge-intensive approach for subgroup discovery utilizing several types of background knowledge, which can be applied incrementally. Our application area is the medical domain of sonography. The context of our work is to identify interesting diagnostic patterns using subgroup discovery techniques, to supplement a medical documentation and consultation system. We present an experimental evaluation of our approach using a case base from a real world application.

1 Introduction

Knowledge discovery in databases (KDD) [Fayyad *et al.*, 1996] is concerned with the computer-aided extraction of novel, potentially useful, and interesting knowledge from (large) databases. *Subgroup discovery* [Wrobel, 1997; Klösgen, 2002] is a subclass of knowledge discovery tasks to discover "interesting" subgroups of individuals. These "interesting" subgroup individuals can be defined as a subset of the target population with a distributional unusualness concerning a certain property we are interested in. Subgroup discovery methods take relations between independent (explaining) variables and a dependent (target) variable into account. These relations are rated by a certain user-defined "interestingness" measure.

The main application areas of subgroup discovery are exploration and descriptive induction, when the user wants to obtain an overview of the dependencies between a specific target variable and usually many explaining variables. Therefore, the subgroup discovery approach does not necessarily focus on finding complete relations between the target and the explaining variables; partial relations, i.e., subgroups with "interesting" characteristics, e.g., with a significant deviation from the total population, 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].

Our application area for subgroup discovery is the medical domain of sonography. In general, in the medical domain often the relations with high support are already known. However, the manual discovery of significant correlations in a restricted population subset is quite difficult. In such a scenario, (automatic) subgroup discovery methods can be applied for descriptive and exploratory data mining, to acquire novel, potentially useful, and interesting knowledge in medical case bases.

Background knowledge can help to improve subgroup discovery in several ways. For example, it can focus the mining algorithm on the relevant patterns according to specific criteria, thus reducing uninteresting patterns and restricting the search space. This helps to improve the quality of the discovered set of subgroups, and also increases the efficiency of the search method. In the medical domain, for example, often a lot of background knowledge is available. This can be utilized by formalizing it and supplying it to the data mining method in a semi-automatic approach.

Besides constraints the applicable background knowledge for the knowledge-intensive process consists of two categories: a task-specific subset of derived attributes used for analysis and general ontological background knowledge. In the knowledge-intensive process for subgroup discovery we basically want to use as much background knowledge as possible. In addition, subgroup discovery results can be formalized as background knowledge incrementally and can be provided to the search-method for further analysis. We will introduce this approach in this paper.

Our implementation and evaluation is based on the knowledge-based documentation and consultation system for sonography SONOCONSULT [Huettig *et al.*, 2004] developed with the diagnostic shell kit D3 [Puppe, 1998]. SONOCONSULT is in routine use in the DRK-hospital in Berlin/Köpenick. The system documents an average of 300 cases per month. In addition to a documentation system, SONOCONSULT also infers diagnoses with heuristic expert knowledge. The cases are detailed descriptions of findings of the examination(s), together with the inferred diagnoses (concepts). The inferred diagnoses can be corrected manually, but are usually correct due to first evaluations of SONOCONSULT. This setting yields a high quality of the case base with detailed and usually correct case descriptions.

The rest of the paper is organized as follows: In Section 2 we describe the knowledge-intensive process for subgroup discovery. We introduce the necessary background knowledge and present the application of background knowledge for subgroup discovery. An experimental evaluation of the knowledge-intensive subgroup discovery method is given in Section 3. We conclude the paper in Section 4 with a discussion of the presented work and we show promising directions for future work.

2 Knowledge-Intensive Subgroup Discovery

In this section, we first give an overview of the proposed knowledge-intensive subgroup discovery process. Then, we describe the basic components of the process in detail. We define the subgroup discovery task and the necessary basic concepts of our knowledge representation schema. After that, we introduce the background knowledge which is necessary for the knowledge-intensive subgroup discovery task. Then, we present the subgroup discovery process utilizing the described background knowledge in a semi-automatic manner.

2.1 The Knowledge-Intensive Process for Subgroup Discovery

Subgroup discovery [Wrobel, 1997; Klösgen, 2002] is a method to identify "interesting" subgroups of individuals. These are defined as a subset of the target population with a distributional unusualness concerning a certain target property. Subgroup discovery takes relations between independent (explaining) variables and a dependent (target) variable into account. These relations are rated by a certain "interestingness" measure. For example, subgroups can be considered, where the distribution of the target variable differs significantly from the general population, and where the subgroups should be as large as possible.

To guide the subgroup discovery process we propose to apply as much background knowledge as possible to support the discovery method. In knowledge-rich domains, e.g., in the medical domain, often a lot of knowledge is available beforehand. Providing this background knowledge to the discovery method can improve the results significantly concerning the interests of the user. Since often the main aim of the mining process is to find novel knowledge, the number of uninteresting results should be reduced. Also, the search space can be constrained significantly. Therefore we favor a knowledge-intensive approach in which background knowledge can be applied initially at the start of the process, but also incrementally during the discovery process.

Types of Background Knowledge There are different classes of background knowledge which are used in the knowledge-intensive process for subgroup discovery. In the following we summarize the main ideas and describe the relation to the knowledge-intensive process.

Constraints are a form of background knowledge, which are simple to apply in the subgroup discovery process:

- Constraints for value-sets of attributes: the attribute values can be restricted to the relevant values, i.e. values can be excluded. Additionally specific attribute groups defining an abstracted value can be specified, e.g., intervals for numerical values. Attribute groups are not restricted to intervals, but can cover any combinations of values of an attribute value range.
- Constraints for attributes: attributes can be excluded from the search space. Furthermore, inclusion and/or exclusion conditions for combinations of attributes can be defined.
- General constraints customizing the search process: constraints restricting the syntactical form of the discovered subgroups, or quality constraints for the discovered subgroups can be applied.

Using such constraints we can both restrict the search space and focus the search process.

Besides constraints there are other types of background knowledge which can be included in the knowledge-intensive process for subgroup discovery:

- Pattern knowledge, which defines already known subgroups, for example. These subgroups can then be directly applied in the discovery process, e.g., for subgroup refinement.
- Ontological knowledge which is typically available in knowledge systems, e.g., attribute weights, similarities between attribute values, and abnormality knowledge about attribute values.
- Abstraction knowledge: this type of background knowledge specifies special ontological knowledge which is constructed according to user requirements and analysis goals.

Using these types of background knowledge, additional constraints can be defined as well. We will discuss the background knowledge in Section 2.3 in more detail.

Process for Knowledge-Intensive Subgroup Discovery The proposed knowledge-intensive process for subgroup discovery is depicted in Figure 1. We start with a defined population given as a case base CB and optionally with existing background knowledge. For the analysis task defined by a *subgroup discovery problem* the subgroup discovery method generates a set of subgroups. If these subgroups are interesting according to the user's goals the results are presented, and the process is finished. Otherwise, the subgroups are analyzed either automatically or semi-automatically guided by user interaction. As a result of this analysis background knowledge and additional constraints for the subgroup discovery problem are provided to the search method. Additionally, selected subgroups can be provided to the subgroup discovery method for refinement. Then the process continues with a new iteration.

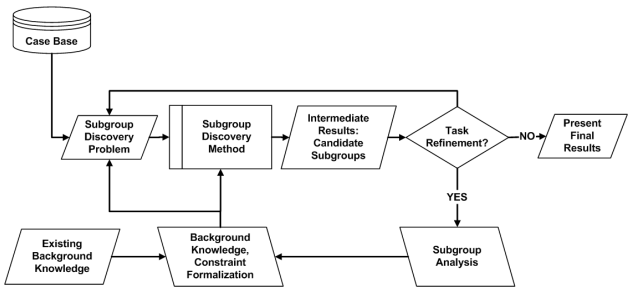

Figure 1: Knowledge-Intensive Process for Subgroup Discovery

2.2 Subgroup Discovery

The main application areas of subgroup discovery are exploration and descriptive induction, when the user wants to get an overview of the dependencies between a certain target variable and usually many explaining variables. The deviations of a subgroup from the performance of the general population are usually not simply due to statistical fluctuations, but are caused by local factors. Identifying these factors helps to understand the data in general and has a huge impact, e.g., on diagnostic, preventive or therapeutic issues concerning medical questions. Thus, subgroup discovery has become more important in the medical domain [Gamberger *et al.*, 2003], lately.

The subgroup discovery task relies on the following four main properties:

- The type of the target variable, i.e., the target analysis object. A target variable may be binary, nominal or numeric. Depending on the type of the target variable different analytic questions are possible. For example, for a numeric target variable we can search for significant deviations of the mean of the target variable.
- The description language specifying the individuals from the reference population belonging to the subgroup. Mainly conjunctive languages are used. The subgroup description consists of a set of selection expressions (selectors). In the simplest case, a selection expression is one-valued, however negation or internal disjunctions are possible, too.
- The quality function measuring the interestingness of the subgroup. A variety of quality functions were proposed (e.g., [Klösgen, 1996; 2002; Gamberger *et al.*, 2003]). The applicable set of quality functions is determined by the type of the target variable and the analytic problem.
- Finally, the search strategy is very important, since the search space is exponential concerning all the possible selectors of a subgroup description. Commonly, a beam search strategy is used due to its efficiency [Klösgen, 2002].

Before introducing the subgroup discovery task methodology, we first define the necessary notions concerning our knowledge representation schema.

Basic Definitions Let Ω_D be the set of all diagnoses and 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 Ω_F to be the (universal) set of findings of the form $(a = v)$, where $a \in \Omega_A$ is an attribute and $v \in dom(a)$ is an assignable value. For each diagnosis $d \in \Omega_D$ we define a range $dom(d)$: $\forall d \in \Omega_D : dom(d) = \{established, not\ established\}$, i.e., the diagnosis denotes a boolean variable.

Let CB be the case base containing all available cases. A case $c \in CB$ is defined as a tuple $c = (\mathcal{F}_c, \mathcal{D}_c)$, where $\mathcal{F}_c \subseteq \Omega_F$ is the set of findings observed in the case c. The set $\mathcal{D}_c \subseteq \Omega_D$ is the set of diagnoses describing the *solution* for this case. The occurrence of a diagnosis d in a case c, i.e., $d \in \mathcal{F}_c, d \in \Omega_D, c \in CB$ indicates the 'finding' $(d = established)$. The value $not\ established$ does not occur in our case base. Thus, the union $\Omega_F \cup \Omega_D$ denotes the (universal) set of all possible generalized findings which can occur in the case base CB.

We define the subgroup description language and the specification of the target property as follows. For the subgroup descriptions, selection expressions (selectors) are used to characterize the subgroup instances.

Definition 1 (Subgroup Description) *As a description of the subgroup instances, a subgroup description $sd = \{e_i\}$ consists of a set of selection expressions $e_i \in \Omega_E$ which are selections on domains of attributes, i.e., $e_i = (a_i, V_i)$, where $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. Ω_E is defined as the set of all possible selection expressions. 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.*

A subgroup discovery problem encapsulates the target property (target variable) for the subgroup discovery task, the search space of independent variables, the general population, and additional constraints.

Definition 2 (Subgroup Discovery Problem) *A subgroup discovery problem SP is defined as the tuple*

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

where $T \in \Omega_A \cup \Omega_F \cup \Omega_D$ is a target variable. $A \subseteq \Omega_A \cup \Omega_D$ 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 discovery. C specifies (optional) constraints for the discovery method. For example, C can contain constraints concerning the construction of selection expressions, i.e., restrictions to sub-domains of attributes contained in A. Furthermore, C can contain additional constraints for the subgroup discovery task, for example range restrictions for the number of variables to include in the subgroup description. We define Ω_{SP} as the set of all possible subgroup discovery problems.

The definition above allows arbitrary target variables. However, for our analytic questions we will focus on subgroup discovery problems with binary target variables, i.e., $T \in \Omega_F \cup \Omega_D$.

Given a subgroup discovery problem $SP \in \Omega_{SP}$, the subgroup search method is guided by a quality function for identifying interesting subgroups.

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

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

evaluates a subgroup description $sd \in \Omega_{sd}$ concerning a specific subgroup discovery problem $SP \in \Omega_{SP}$. It is used by the search method to rank the discovered subgroups when processing the defined search space.

The applicable quality functions are dependent on the type of the target variable. Since we restrict our analytic questions to binary target variables, e.g., a patient has a certain disease or not, we present two exemplary quality functions for this kind of questions.

A classic quality function described by [Klösgen, 1996] is the binomial test.

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

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. This test takes both the deviation of the subgroup from the total population and the size of the subgroup into account.

Another quality function which is especially useful in the medical context is described by [Gamberger *et al.*, 2003].

$$q_{TP} = \frac{tp}{fp + g}\,, \qquad (2)$$

where tp denotes the number of *true positives*, i.e., the number of objects belonging to the subgroup that also contain the target variable. fp denotes the remaining elements of the subgroup, i.e. the *false positives*. g is a generalization parameter. If g is set to low values, then fewer false positives are tolerated. Otherwise more general subgroups are allowed.

In contrast to the function q_{BT}, the quality function q_{TP} does not measure the deviation from a reference population, but assigns the highest rank to subgroups with a maximum number of true positives and a minimum number of false positives.

For our search strategy, we use a modified beam search strategy, where an initial subgroup description can be selected as the initial value for the beam. Beam search starts with a subgroup discovery problem $SP \in \Omega_{SP}$ and a given initial subgroup description $sd \in \Omega_{sd}$. If no initial subgroup description is provided, then sd does not contain any selectors. In each iteration a selection expression is added to the subgroup description. Then, the quality of the new subgroup description is evaluated by a quality function q using the subgroup discovery problem SP. For each beam search iteration the k best subgroup descriptions are used in the next iteration until the quality of the k best subgroup descriptions is not improved any further.

For 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 generalized findings $supp \subseteq \Omega_F \cup \Omega_D$ contained in the subgroup, which are characteristic for the subgroup, i.e., the value distributions of their corresponding 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 in [Gamberger and Lavrac, 2002], for example.

2.3 Integrating Background Knowledge Into the Subgroup Discovery Process

For the knowledge-intensive process for subgroup discovery, different types of background knowledge can be applied. Constraints can be used to define the search space, i.e., the space of attributes and attribute values used in the search process can be restricted. Furthermore relations between attributes and attribute values which should be enforced during the search process can be specified. Additionally, constraints can specify characteristics of the search process, i.e., restricting the pattern language to enforce simplicity constraints, for example.

Besides the conceptually simple class of constraints we propose general ontological knowledge and abstraction knowledge as suitable background knowledge for the knowledge-intensive process. Based upon these types of background knowledge, we can additionally form new constraint knowledge as described below.

Ontological Background Knowledge

For the first class of background knowledge we can utilize general forms of ontological knowledge which are commonly known in the development of knowledge systems, e.g., in CBR systems. The following elements can initially be defined by an expert, or can be automatically learned (e.g. [Baumeister *et al.*, 2002]).

- weights of attributes, which denote the importance of attributes,
- similarity information about the relative similarity between attribute values,
- abnormality information about attribute values

Weights of attributes provide an easily applicable form of feature subset selection. Attributes can be included into the subgroup discovery process depending on their weights, which denote their relative importance. Thus, by applying knowledge about weights of attributes the large search space can be reduced.

Likewise, abnormality information about attribute values can be used to constrain the value range of an attribute used in the subgroup discovery method. If *abnormality* information about attribute values is available, then each value $v \in dom(a)$ of an attribute a is attached with a label that explains, if v is describing a normal or an abnormal state of the attribute. 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. Several categories can be defined according to the degree of abnormality. Up to now, we use five degrees of abnormality, i.e. given by the symbolic values $\{A1, A2, A3, A4, A5\}$. Category $A1$ denotes a normal value. The remaining categories $\{A2, A3, A4, A5\}$ denote abnormal values in ascending order.

Furthermore, abnormality information and similarity information concerning attribute values can be used to define additional *abstracted* attribute values, i.e., constraints on special attribute values: if the similarity between two attribute values is very high, then they can potentially be analyzed as one value, thus forming a disjunctive selection expression on the value range of an attribute. Likewise, abnormality groups can be defined using the abnormality categories. The user can define groups of abnormality degrees which specify, which values should be included into a disjunctive selection expression. This is especially relevant in medical domains where attributes can have values such as *probable*, *possible*, and *unverifiable*. In diagnostics the value *probable* contributes more evidence to the concept represented by the attribute than the value *possible*. Therefore essentially only *probable* should be analyzed. However, often the values *probable* and *possible* can be analyzed together to enable greater support for hypothesis testing.

Depending on the specific analytic question, values can be excluded from the range of an attribute used in the subgroup discovery method, by an exclusion constraint. The criterion for exclusion is given by the abnormality of a value. For example, either *normal* (non-interesting) values and/or selected abnormal values can be excluded. Thus, abnormality information is used for constraining the range of values of an attribute in the subgroup search method. This restricts the set of possible selection expressions which can be constructed for the attribute.

Applying the class of ontological background knowledge we can add a set of constraints to the constraints C of the subgroup discovery problem $SP \in \Omega_{SP}$. The set of relevant attributes can be constrained using attribute weights. Using similarity and abnormality information of attribute values we can both model and restrict the value ranges of attributes, as described above.

Abstraction Knowledge

The second type of background knowledge is given by derived attributes, which are constructed especially for subgroup discovery purposes. These attributes are inferred from basic attributes or other derived attributes. The derived attributes can be constructed by the expert before performing subgroup discovery, and often correspond to certain known dependencies between attributes. For exam-

ple, in the medical domain derived attributes can denote common intermediate concepts, that are not stored in the case base. Then, a derived attribute concept can be constructed quite easily. For example, if we consider the derived attributes *pleura-effusion - left*, and *pleura-effusion - right*, then we can infer the general derived attribute *pleura-effusion - sonographic*, from both.

Additionally, if there are a lot of basic attributes in the case base, then the huge number of analysis objects may cause unstructured subgroup discovery results because of possible multi-correlations between basic attributes. In this case, data abstraction can be very important. It can increase the interpretability of the knowledge discovery results significantly, because simple concepts can be aggregated to intermediate concepts to form more potentially meaningful, interesting, and significant selectors.

A nominal derived attribute $a \in \Omega_A$ is defined using abstraction rules, which are utilized to derive the findings $f_{i_a} \in \Omega_F$ concerning attribute a. A rule of the form

$$r_{f_a} = cond(r_{f_a}) \rightarrow f_a \,,$$

is used for a finding f_a of attribute a, where the rule condition $cond(r_{f_a})$ contains conjunctions and/or disjunctions of (negated) generalized findings $f_i \in \Omega_F \cup \Omega_D$. Furthermore, derived attributes with a numerical value range can be defined by algebraic formula expressions.

Improving the Handling of Missing Values

Considering the quality functions abstraction knowledge contributes to one major point – handling missing values. Missing values in cases are a significant problem for knowledge discovery in medical case bases. For a specific patient only a subset of the possible examinations is usually performed, which results in many missing attribute values. The documentation and consultation system SONOCONSULT is a knowledge system guiding the data acquisition process by rules. Only the relevant questions for the diagnostic tasks are presented to the user. This results in reduced effort for the examiner, however then a specific instance of the data set concerning the basic attributes may be quite sparse.

There exist several strategies for dealing with missing values: the standard strategy [Tsumoto, 2002] removes objects (cases) with missing values from the set of analysed objects. Other strategies try to fill in the missing values according to statistical evaluations, or try to model the distribution [Ragel and Créwmilleux, 1999]. The quality functions basically perform a form of statistical hypothesis testing given a subgroup description, the target variable and the general population. For such a test only the cases of the population can be considered in which all variables have defined values. The power of the test is decreased significantly if a lot of analysis objects are removed due to missing values.

In the medical domain we cannot simply apply the "closed-world assumption", i.e., that missing values of a concept indicate the non-existence or negation of the concept. For example, a diagnosis may be missing, because either all its relevant observations are missing or they are known but denote the normal, i.e., then non-pathological state. In effect the diagnosis is not inferred. If we construct a derived attribute to indicate the cases when the relevant observations are missing, then we can use this derived "helper" attribute as a filter: we remove the cases in which the relevant observations are missing, and apply the closed-world assumption for the remaining cases.

Additionally, derived attributes besides the described "helper" attributes can also be constructed accordingly to minimize missing values themselves, such that a certain default value is provided which denotes the normal category. So, the derived attributes serve three purposes:

- they focus the subgroup discovery method on the relevant analysis objects,
- they decrease multi-correlations between attributes that are not interesting,
- derived attributes can minimize missing values for a given concept, since they can be constructed such that a defined value is more often computed if the respective concept would have a missing value otherwise.

The derived attributes can either be constructed based on expert knowledge, or on subgroup discovery results, i.e., (sets of) subgroups. Subgroup discovery results in a set of selectors for a specific target concept that are highly correlated with the concept. If the selectors can be abstracted into a derived attribute, then this attribute can be used as potential background knowledge as well. Furthermore, derived attributes can be refined according to the subgroup discovery results. The attributes can be specialized by including more selection expressions into their derivation process. In contrast, they can also be generalized by removing redundant or irrelevant attributes without a significant correlation, which were included erroneously.

2.4 Related Work

There exist several approaches for subgroup discovery. In general, these can be divided into purely automatic methods, and methods which integrate user constraints. The *EXPLORA* system [Klösgen, 1996] offers various search-strategies for general automatic subgroup discovery tasks. The system is also able to integrate simple constraints, i.e., taxonomies of attribute values, which are similar to value groups.

A special adaptation of how to use a standard rule-learning algorithm for subgroup discovery is described in [Lavrac *et al.*, 2004]. This approach is also a purely automatic approach without user-interaction.

[Wrobel, 1997] proposes a method for multi-relational subgroup discovery implemented in the *MIDOS* algorithm. Another system which uses a multi-relational hypothesis space is the *SubgroupMiner* system [Klösgen and May, 2002] for spatial subgroup mining. In addition, SubgroupMiner also supports causal analysis on the discovered set of subgroups.

The application of subgroup discovery especially for the medical domain using the expert's guidance is described in [Gamberger and Lavrac, 2002; Gamberger *et al.*, 2003]. This approach stresses the interaction between the expert and the system to identify interesting subgroups. Also, the analysis task differs slightly from the approaches described above, since a new quality function is introduced which is especially well suited for medical subgroup analysis. However, in the semi-automatic process only the parameters of the search process can be adapted.

For our approach multi-relational subgroup discovery is not necessary, since the case base is given in one relation. Likewise, causal analysis is not a major priority so far, because the analysis objects, i.e., the derived attributes are specifically constructed to the question at hand. In the construction process multi-correlations between attributes should be taken into account, as far as possible.

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 groups of literals, i.e., 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 restrictions and constraints relating to the quality of the discovered subgroups.

The main difference between our approach and the existing approaches is the fact, that we are able to integrate several new types of additional background knowledge. This additional background knowledge can be refined incrementally according to the requirements of the discovery task, and can additionally be used quite easily to infer new background knowledge on the fly, e.g., constraints.

As the major point we apply special abstraction knowledge, which can be defined by the expert, or can be constructed semi-automatically using the subgroup discovery results. This type of knowledge can be applied dynamically in the process and does not rely on static data-preprocessing and cleaning task, for example. Then, in a semi-automatic manner the user/expert can inspect the results of the subgroup discovery process to modify the subgroup discovery problem, to include additional constraints, or to modify the available background knowledge.

3 Experimental Evaluation

We evaluated the presented approach with cases taken from the medical application SONOCONSULT, which is currently in routine use. The applied SONOCONSULT case base contains 4358 cases. The domain ontology of SONOCONSULT 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. This indicates the huge search space formed of all possible attributes for subgroup discovery. In the following, we describe experiences conducting our approach in an experimental evaluation. We used beam search with a beam size of 10 as the search strategy, and the standard binomial test quality function defined in Equation 1. The discovered subgroups were evaluated by a medical expert of the application domain. The expert assessed which of the subgroups were new, interesting, and thus appropriate for clinical practice.

First, we performed subgroup discovery only using basic attributes and general background knowledge. We used attribute weights for feature subset selection. The subgroup discovery algorithm presented many significant subgroups. However, these subgroups mostly indicated dependencies that were already known to the expert, and were already formalized as diagnostic knowledge contained in the SONOCONSULT knowledge base. These results supported the applicability of the subgroup discovery techniques for the domain, but the results were not really interesting for the expert concerning the novelty aspect.

Therefore, the expert decided to define new attributes, i.e., abstracted attributes which described interesting concepts for the analysis. The expert provided 45 derived attributes, which were constructed to minimize missing values. Parts of the derived attributes are symptom interpretations which directly indicate a diagnosis. The rest of the derived attributes denote intermediate concepts which are used in clinical practice, for example *pleura-effusion*, or *portal hypertension*.

In the next stage, the newly defined abstraction knowledge was applied extending the search space to the expert-defined attributes. For each attribute a in the set of derived attributes and each value $v_i \in dom(a)$, a subgroup discovery problem $SP_{ai} \in \Omega_{SP}$ was generated; the target variable was given by the binary target variable $(a = v_i)$. Then the subgroup discovery algorithm was applied on the defined subgroup discovery problems. The impact of the added background knowledge was proven by a greater acceptance of the subgroup discovery results by the expert. The resulting subgroups were often significant at the 0.05 level, however many subgroups contained selection expressions which were not interesting for the expert.

This was due to the fact, that too many *normal* values were included in the results, which motivated the application of abnormality information to constrain the value space to the set of *abnormal* values of the attributes. Additionally, the expert suggested to group sets of values into disjunctive value sets defined by abnormality groups. For example, we extended the value range of selected attributes such that the values *probable* and *possible* are considered as a new disjunctive value.

After applying this background knowledge the results were regarded as potentially interesting for clinical practice. Further investigation showed that missing values play a central role in the discovery process. Sometimes the defined population significantly decreased, when adding a selection expression to a subgroup description, compared to the parent subgroup. Thus, the respective derived attributes were adapted accordingly. After that, the final sets of subgroups for the subgroup discovery problems were obtained.

To assess the discovered subgroups concerning interestingness for clinical practice, the expert wanted to obtain a quick overview of interesting subgroups for a first estimation. Therefore, we applied a rating function q_{RG} similar to the binomial test quality function for result analysis. This functions was applied on the discovered subgroups, i.e., to post-process these and to present relevant subgroups to the expert. The function q_{RG} is defined as follows:

$$q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)}\,, \tag{3}$$

where p is the relative frequency of the target variable in the subgroup and p_0 is the relative frequency of the target variable in the total population. This interestingness function measures the relative gain of the probability of the target variable in the subgroup compared to the total population. Then, suitable gain thresholds can be used helping the expert to identify interesting subgroups. As a general principle, the expert preferred smaller subgroup descriptions, which is in line with the heuristic of preferring simpler knowledge for actionability.

For the evaluation the expert selected 40 subgroups as especially interesting from the total number of 605 discovered subgroups. The q_{RG} values measuring the relative gain there range from 1.1 to 20.8, i.e. from a 110% gain to a 2080% gain. The selected subgroups with a high relative gain are sometimes also quite large subgroups. This is in contrast to the results considering the set of all discovered subgroups; a maximum gain value of 96 was achieved, however only for a small subgroup. This is not too surprising, because the binomial test quality function additionally takes the subgroup size into account. However, for our analysis the expert chose the q_{RG} measure as an easy to interpret measure for post-processing and comparing the discovered individual subgroups.

In the following table we show an example of three significant subgroups, where the most special one is specialized on the two more general ones. The subgroups were significant at the 0.0005 level. Column q_{RG} shows the relative gain measure. In this example specializing the subgroup significantly increased the subgroup quality compared to the parent subgroups.

Target Variable	q_{RG}	*Subgroup Description*
SI-fatty liver = probable	0.026	SI-liver size = marginally increased
SI-fatty liver = probable	0.111	SI-aorto-sclerosis = not calcified
SI-fatty liver = probable	3.48	SI-liver size = marginally increased AND SI-aorto-sclerosis = not calcified

4 Summary and Future Work

In this paper we presented a knowledge-intensive approach for subgroup discovery. For the knowledge-intensive process we discussed applicable background knowledge in more detail. We described how the application of abstraction knowledge can help to handle the problem of missing values, which is often experienced in medical case bases. An experimental evaluation performed by a domain expert showed that applying background knowledge helped to focus the discovery algorithm on the interesting subspace of subgroup hypotheses.

In the future we are planning to consider appropriate quality measures concerning the simplicity of the discovered subgroups. Primary work for learned rule bases was presented in [Atzmueller *et al.*, 2004]. As a related direction, we will further focus on quality measures which are especially easy to interpret for the expert, and tuneable to the analysis goals of the expert. Furthermore, we will investigate the application of automatic construction methods for derived attributes in order to support the expert in the semi-automatic process. Additionally, the impact of causal analysis in subgroup discovery is an interesting issue to consider in the future.

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.*, 2004] Martin Atzmueller, Joachim Baumeister, and Frank Puppe. Quality Measures for Semi-Automatic Learning of Simple Diagnostic Rule Bases. In *Proc. 15th International Conference on Applications of Declarative Programming and Knowledge Management (INAP 2004)*, pages 203–213, Potsdam, Germany, 2004.

[Baumeister *et al.*, 2002] Joachim Baumeister, Martin Atzmueller, and Frank Puppe. Inductive Learning for Case-Based Diagnosis with Multiple Faults. In *Advances in Case-Based Reasoning*, volume 2416 of *LNAI*, pages 28–42, Berlin, 2002. Springer Verlag. Proc. 6th European Conference on Case-Based Reasoning.

[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.

[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 and May, 2002] Willi Klösgen and Michael May. Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proc. Principles of Data Mining and Knowledge Discovery. 6th European Conference, PKDD 2002*, volume 2431 of *LNCS*, pages 275–286, Berlin, 2002. Springer Verlag.

[Klösgen, 1996] Willi Klösgen. Explora: A Multipattern and Multistrategy Discovery Assistant. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 249–271. AAAI Press, 1996.

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

[Lavrac *et al.*, 2004] Nada Lavrac, Branko Kavsek, Peter Flach, and Ljupco Todorovski. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

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

[Ragel and Créwmilleux, 1999] A. Ragel and B. Créwmilleux. MVC–a Preprocessing Method to Deal with Missing Values. *Knowledge-Based Systems*, 12(5–6):285–291, October 1999.

[Tsumoto, 2002] Shusaku Tsumoto. *Handbook of Data Mining and Knowledge Discovery*, chapter Medicine. Chapter 43. Oxford University Press, New York, 2002.

[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 Jan Komorowski and Jan Zytkow, editors, *Proc. First European Symposion on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, 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.