Difference-based Estimates for Generalization-Aware Subgroup Discovery

Florian Lemmerich, Martin Becker, and Frank Puppe

Artificial Intelligence and Applied Computer Science Group,
University of Würzburg, D-97074 Würzburg, Germany
{lemmerich, becker, puppe}@informatik.uni-wuerzburg.de

Abstract. For the task of subgroup discovery, generalization-aware interesting measures that are based not only on the statistics of the patterns itself, but also on the statistics of their generalizations have recently been shown to be essential. A key technique to increase runtime performance of subgroup discovery algorithms is the application of optimistic estimates to limit the search space size. These are upper bounds for the interestingness that any specialization of the currently evaluated pattern may have. Until now these estimates are based on the anti-monotonicity of instances, which are covered by the current pattern. This neglects important properties of generalizations. Therefore, we present in this paper a new scheme of deriving optimistic estimates for generalization aware subgroup discovery, which is based on the instances by which patterns differ in comparison to their generalizations. We show, how this technique can be applied for the most popular interestingness measures for binary as well as for numeric target concepts. The novel bounds are incorporated in an efficient algorithm, which outperforms previous methods by up to an order of magnitude.

1 Introduction

Subgroup discovery [17] is a key technique for data mining and machine learning. It aims at identifying descriptions for subsets of instances in a dataset, which have an interesting deviation with respect to the distribution of a predefined concept of interest. This task has been studied under different terminology such as contrast set mining [7], emerging pattern mining [11], correlated itemset mining [21], discriminative pattern mining [10] or association rule mining with a fixed consequent [20]. While the specific goal of these tasks may vary, the algorithmic challenges and approaches are very closely related, see [18,26].

The selection of patterns in the search space is commonly based on an interestingness measure. These measures use statistics derived from the instances covered by a pattern to determine a score for the pattern. The best patterns according to this score are then returned to the user. As an example, consider a dataset of patients and their medical data. Let the target concept be `surgery successful`, which is true for 30% of the patients. Then a pattern like `gender=male` $\wedge$ `smoker=false` with a higher rate of successful surgeries, e.g. 50%, receives a higher score and is more likely to be included in the result.

Practical applications have shown that results for traditional interestingness measures often contain variants of the same pattern multiple times. To avoid this problem, several authors postulated that a pattern should not only be evaluated with respect to its own statistics, but also with respect to the statistics of its generalizations, see for instance [8,4,5,19]. Considering the example above the pattern `gender=male` $\wedge$ `smoker=false` would be rated as less interesting if it can be explained by one of its generalizations alone, e.g., if the pattern `smoker=false` already describes a set of patients with a 50% surgery success rate. While the practical use of such generalization-aware interestingness measures has been widely acknowledged, the efficient mining in this setting has received little attention. A key technique to improve runtime performance of subgroup discovery in general is the application of optimistic estimates, that is, upper bounds for the interestingness of any specialization of the currently evaluated pattern. Although research has shown that improving the tightness of the utilized bounds improves the runtime performance substantially [15], there has been no extensive research so far concerning upper bounds for generalization aware interestingness measures beyond the trivial transfer of bounds for traditional measures.

In this paper we propose a novel method to exploit specific properties of generalization-aware measures to derive additional optimistic estimate bounds, which allow to speed-up the search. Unlike previous approaches, the bounds are not exclusively based on the instances that are contained in the currently examined subgroup, but on the instances that were excluded in comparison to generalizations of the current pattern. We show, how this general concept can be applied to exemplary interestingness measures in different setting, i.e., for subgroup discovery with binary target concepts and with numeric target concepts using a mean-based interestingness measure. The bounds are incorporated in a novel *apriori*-based algorithm that allows efficient propagation of the required statistics. Experiments show that exploiting the presented bounds results in substantial runtime improvements. The optimistic estimates are especially effective in tasks that incorporate selectors, which cover a majority of the dataset.

The rest of the paper is structured as follows: Section 2 provides background on subgroup discovery and the used terminology. Then, related work is discussed in Section 3. Next, the new scheme to derive optimistic estimate bounds and its application to different interestingness measures is presented in Section 4. Afterwards, we explain, how the new optimistic estimate bounds can efficiently be exploited in an algorithm in Section 5. Section 6 presents experimental results, before we conclude in Section 7.

2 Background

Let A be an *attribute space* $A = A_1 \times \ldots \times A_m$, where each set A_i represents an *attribute*. A *dataset* is a tuple $D = (I, A)$ with $I \subseteq A$. Each $i \in I$ is called a *data instance*. *Selectors sel* (also called basic patterns) are boolean functions $sel : I \rightarrow \{false, true\}$ defined by selection expressions on the set of attributes. In

the case of nominal attributes typical selection expressions are given by attribute-value pairs, in the case of numeric attributes by intervals. For example, the selector $age =]12; \infty[$ is true, iff the attribute *age* has a value greater than 12. A (complex) *pattern* (also called subgroup description) combines selectors into a boolean formula. For a typical conjunctive description language, on which we focus in this paper, a pattern $P = \{\text{sel}_1, \ldots, \text{sel}_k\}$ is defined by a set of selectors sel_j, which are interpreted as a conjunction, i.e., $P = \text{sel}_1 \wedge \ldots \wedge \text{sel}_k$. Thus, an instance $i \in I$ is *covered* by pattern P, iff $(\forall sel \in P : sel(i) = true)$ or short $P(i) = true$. A subgroup $sg(P)$ is given by the set of individuals covered by the pattern P: $sg(P) = \{i \in I | P(i) = true\}$. For short notation, $i_P = |sg(P)|$ is the number of individuals covered by a pattern P. Furthermore we denote as $\Delta(A, B) = sg(A) \setminus sg(B)$ the instances, which are covered by A, but not by B. We call a pattern G a *generalization* of its *specialization* S, iff $G \subset S$.

A *pattern mining task* is specified by a 5-tuple (D, T, q, Σ, k). D is a dataset. The target concept T assigns a target value $tc(i)$ to each instance. It can either be defined by a pattern (binary case) or by a single numeric attribute (numeric case). In the binary case, we write p_P (n_P) for all individuals with a true (false) target concept. $q : 2^\Sigma \to \mathbb{R}$ is a quality function that measures the interestingness of a pattern with respect to the chosen target concept T. Σ defines the search space by providing a set of selectors to build conjunctive patterns from. k specifies the number of patterns contained in the result set. The overall task is then to identify the best k patterns in the search space 2^Σ according to the quality function q ($q \gg 0$).

A huge amount of quality functions has been proposed in literature, cf. [17,13]. While the general approach of this paper could also be applied to other quality functions, we especially focus on the following popular measures: The most popular interestingness measures trade-off the covered instances i_P of a pattern versus the deviation of the target share $\tau_P - \tau_\emptyset$, where $\tau_P = \frac{p_P}{p_P + n_P}$ is the ratio of positive instances versus all instances in pattern P and $\tau_\emptyset$ is the same ratio for the overall population. This is formalized as:

$$q^a_{bin}(P) = i^a_P \cdot (\tau_P - \tau_0), a \in [0; 1]$$

This includes for example the weighted relative accuracy for the size parameter $a = 1$, a simplified binomial function for $a = 0.5$, or the added value for $a = 0$. For numeric target concepts this can easily adapted by replacing the target share for the pattern and the overall population with the respective mean values μ_P and $\mu_\emptyset$ of the target attribute:

$$q^a_{num}(P) = i^a_P \cdot (\mu_P - \mu_\emptyset), a \in [0; 1].$$

This definition includes the mean test quality function [17] for $a = 0.5$ and the impact quality function [23] for $a = 1$.

Consider a pattern P with an interestingly high target share τ_P. If another selector *sel* with $sel \notin P$ is added to the pattern P, which does either cover a majority of the instances of P ($sg(P) \approx sg(P \wedge sel)$) or is statistically independent from P and the target concept, then the pattern $P \wedge sel$ will have roughly the same target share as pattern P. Thus, this pattern may also receive a high score according to the previously presented quality measures due to its high

target share. However, it should not be presented to users in the result set, since the additional selector *sel* does not contribute to the increased target share. To avoid such redundant output the *minimum improvement constraint* has been introduced, see [8]. By using this additional filter, all patterns with a target share that is lower or equal to the target share of any of its generalizations are removed from the result set. Nonetheless, patterns that improve the target share only by a small margin, e.g., due to noise, will still be contained in the result. Therefore, more recent approaches incorporate the comparison of pattern statistics with the statistics of its generalizations directly into the interestingness measure [5,14,19] resulting in *generalization-aware interesting measures*. The target share (or the mean value in case of numeric targets) within the pattern is not compared to the target share (mean value) of the overall population but to the maximum target share (mean value) of all its generalizations:

$$r^a_{bin}(P) = i^a_P \cdot (\tau_P - \max_{H \subset P} \tau_H), a \in [0;1]$$

$$r^a_{num}(P) = i^a_P \cdot (\mu_P - \max_{H \subset P} \mu_H), a \in [0;1]$$

Thus, a pattern is only regarded as interesting if its target share (mean value) is considerably higher than it is in all of its generalizations. Although other interestingness measures can be adapted accordingly, we focus on these two families of generalization-aware measures in this paper, since they are the only ones, which have been described in previous literature and applied in practical applications. We will also not argue about advantages of these functions in comparison to traditional measures or other methods that avoid redundant output, such as closed pattern [12], but focus on efficient mining for these generalization-aware measures by introducing novel, difference-based optimistic estimates.

The concept of optimistic estimates has been introduced in order to speed up the subgroup discovery task, see [22,25]. The basic idea of optimistic estimates is as following: if one can guarantee that no specialization of the currently evaluated pattern will have a quality, which is good enough to include the respective pattern into the result set, then we can safely omit these patterns from the search. In doing so we can substantially reduce the number of patterns, which have to be evaluated, while maintaining the optimality of the results. In this regard, we aim at as strict as possible bounds to reduce the remaining search space and thus to speed up the search process. Formally, given a pattern P and an interestingness measure q an optimistic estimate function $oe_q(P)$ is a function such that for each specialization $S \supset P$ of P the quality is lower than the value of the optimistic estimate function for pattern P: $\forall S \supset P : q(S) \leq oe_q(P)$.

3 Related Work

Subgroup discovery is a long studied field [17]. An essential technique for efficient discovery showed to be pruning based on optimistic estimates [22,25]. As Grosskreutz et al. showed, the efficiency of the pruning is strongly influenced by the *tightness* of the bounds [15]. A more general method to derive optimistic estimates for a whole class of interestingness measures, that is, *convex* measures,

was introduced in [20] and later extended in [26]. In this paper, we provide a different technique to determine optimistic estimates to another family of interestingness measures, i.e., generalization-aware measures.

The necessity to consider also generalizations of patterns in selection criteria has been recognized in [8,4]. These early approaches used a *minimum improvement constraint*, which is applied only as a post-processing operation after the mining algorithm. Webb and Zhang presented an efficiency improvement in mining with this constraint in the context of association rules [24] by introducing a pruning condition based on the difference in covering. While the method of Webb and Zhang requires *full* coverage on all instances, the method presented in this work can also be applied with only partial coverage. In addition our method is used to derive upper bounds for interestingness measures instead of exploiting constraints and is also applied in settings with numeric target concepts.

Recent approaches incorporate differences with respect to generalizations directly in the interestingness measure. This showed positive results in descriptive [14,19] as well as predictive settings [6] for both binary and numeric target concepts. However, these papers focus more on which patterns are to be selected and not on efficient mining through pruning. As an exception, Batal and Hausknecht utilized a pruning scheme in an Apriori-based algorithm that is based exclusively on the positives covered by a subgroup [5]. This algorithm is used for comparison in the evaluation section. Utilizing pruning in settings with numeric concepts of interest is more challenging than in the binary case [2]. While for the impact measure q^1_{num} an optimistic estimate has been employed [2,23] in the standard subgroup setting, to the authors knowledge no other pruning bounds for numeric generalization-aware measures have been proposed so far.

4 Estimates for Generalization-Aware Subgroup Mining

In this section, we introduce a novel scheme to derive optimistic estimates for generalization-aware interestingness measures. These optimistic estimates help to improve the runtime performance of algorithms by pruning the search space. We start by generalizing estimates that have been previously presented for this task to outline the conventional approach to derive estimates. Then, we present the core idea of our new scheme to derive upper bounds: difference-based optimistic estimates. Next, we show how this concept can be exploited by deriving estimates for quality functions in the binary and the numeric case using the quality functions r^a_{bin} and r^a_{num}.

4.1 Optimistic Estimates Based on Covered Positive Instances

Traditionally, optimistic estimates for subgroup discovery are based only on the anti-monotonicity of instance coverage. That is, when adding an additional selector to a pattern P, then the resulting pattern only covers a subset of the instances covered by P. To give an example for this traditional approach, the following theorem generalizes the optimistic estimate bounds for r^a_{bin} used in [5], which covers only the special case using the parameter $a = 0.5$.

Theorem 1. *Let p_P be the number of all positive instances covered by the currently evaluated pattern P and $\max_{H \subset P}(\tau_H)$ the maximum of the target shares for P and any of its generalizations. Then, optimistic estimate bounds for the family of quality functions r^a_{bin} are given by: $oe_{r^a_{bin}} = (p_P)^a \cdot (1 - \max_{H \subseteq P} \tau_H)$.*

Proof. We first show that the quality of any specialization S does not decrease, if all negatives are removed. Let n_s be the number negatives in S. Then,
$r^a_{bin}(S) = (p_S + n_S)^a \cdot (\frac{p_S}{p_S + n_S} - \max_{H \subset S} \tau_H) = \frac{p_S}{(p_S + n_S)^{1-a}} - (p_S + n_S)^a \cdot \max_{H \subset S} \tau_H$
We examine this term as a function of n_S, $n_S \geq 0$: The first summand decreases with increasing n_S, since $1 - a \geq 0$. The second, negative summand increases with increasing n_S, as $max_t \geq 0$. Thus, the maximum is reached for $n_S = 0$. We can conclude that:

$$\begin{aligned} r^a_{bin}(S) &= (p_S + n_S)^a \cdot (\frac{p_S}{p_S + n_S} - \max_{H \subset S} \tau_H) \\ &\leq (p_S)^a \cdot (\frac{p_S}{p_S} - \max_{H \subset S} \tau_H) \leq (p_P)^a \cdot (1 - \max_{H' \subseteq P} \tau_{H'}), \end{aligned}$$

as the number of positives in the specialization S is smaller than the number of positives in the more general pattern P, and the generalizations of S include all generalizations of P. □

As has been exemplified in [5] this bound can already achieve significant runtime improvements. Note, that these bounds use only the anti-monotonicity of the covered positive instances. In contrast, we will show in the next sections, how we can exploit additional information on the difference of negative instances between patterns and their generalizations to derive additional bounds.

4.2 Difference-based Pruning

Next, we provide the core idea for our novel scheme to derive optimistic estimates. It utilizes that the instances by which a pattern and its specialization differ are – in a certain way – anti-monotonic. More specifically, we will exploit the following lemma to derive optimistic estimates:

Lemma 1. *Let $P = A \wedge B$ be any pattern with A, B potentially being a conjunction of patterns themselves and $B \neq \emptyset$. Then for any specialization $S \supset P$ there exists a generalization $\gamma(S) \subset S$, such that $\Delta(\gamma(S), S) \subseteq \Delta(A, B)$.*

Proof. Consider for any specialization $S = A \wedge B \wedge X$ (X being potentially a conjunction itself) the pattern $\gamma(S) = A \wedge X$, which is a real generalization of S, since $B \neq \emptyset$. Then, $\Delta(\gamma(S), S) = sg(A \wedge X) \setminus sg(A \wedge B \wedge X) = (sg(A) \cap sg(X)) \setminus (sg(A) \cap sg(B) \cap sg(X)) = sg(X) \cap (sg(A) \setminus (sg(A) \cap sg(B)) = sg(X) \cap (sg(A) \setminus sg(B)) = sg(X) \cap \Delta(A, B)$, which is a subset of $\Delta(A, B)$. □

The subset property implies directly that for each specialization S the generalization $\gamma(S)$ contains at most $i_{sg(S)} + i_{\Delta(A,B)}$ instances. Additionally, in the

case of a binary target, we can estimate the number of negative instances in this generalization: $n_{\gamma(S)} \leq n_S + n_{\Delta(A,B)}$. Furthermore, in the case of a numeric target, the minimum target value of $\Delta(\gamma(S), S)$ is higher than the minimum target value in $\Delta(A, B)$. In mining algorithms, statistics for $\Delta(A, B)$ can be computed with almost no additional effort. For instance, n_A and $n_{A \wedge B}$ are both required anyway in order to evaluate the pattern $A \wedge B$ with r^a_{bin}. Then, $n_{\Delta(A,B)}$ is given by $n_{\Delta(A,B)} = n_A - n_{A \wedge B}$.

As an example, assume that the pattern A covers 20 positive and 10 negative instances and the evaluation of the pattern $A \wedge B$ shows that this pattern also covers 10 negative instances. That is, B covers all negative instances, which are covered by A, $n_{\Delta(A,B)} = 0$. Now consider any specialization S of this pattern. According to the lemma, S has another generalization $\gamma(S)$ that contains the same number of negative instances as S since $n_{\gamma(S)} \leq n_S + n_{\Delta(A,B)}$. As S (as a specialization of $\gamma(S)$) additionally has no more positive instances than S, the target share in S is equal or smaller than for its generalization $\gamma(S)$. Thus, the quality of S according to any generalization-aware measure r^a_{bin} is ≤ 0. Since this is the case for any specialization of $A \wedge B$, specializations of $A \wedge B$ can be pruned from the search space without influencing the results.

This is an extreme example: *all* negative instances of A are also covered by $A \wedge B$. Now assume that $A \wedge B$ had covered only 8 negative instance, thus $n_{\Delta(A,B)} = 10 - 8 = 2$. In this case the lemma guarantees that S has a generalization $\gamma(S)$ with *at most 2* negative instances more than S. If S itself covers a decent amount of instances, the target share in S cannot be much higher than in $\gamma(S)$. Thus, either S is small or there is only a small increase (or a decrease) in the target share comparing S and its generalization $\gamma(S)$. In both cases, the interestingness of S according to r^a_{bin} is low.

Overall we conclude that, if the difference of covered instances between A and $A \wedge B$ is small, then the interestingness score for all specializations is limited. In the next sections we formalize these considerations by deriving formal optimistic estimate bounds that can be used to prune the search space.

4.3 Difference-based Optimistic Estimates for Binary Targets

Following, we provide for generalization-aware measures $r^a_{bin} = i^a_P \cdot (\tau_P - \max_{H \subset P} \tau_H)$ with binary targets new optimistic estimates, which are based on the difference of pattern coverage in comparison to the coverage of generalizations.

Theorem 2. *Consider the pattern P with p_P positive instances. $P' \subseteq P$ is either P itself or one of its generalizations and $P'' \subset P'$ a generalization of P'. Let $n_\Delta = n_{P''} - n_{P'}$ be the difference in coverage of negative instances between these patterns. Then, an optimistic estimate of P for r^a_{bin} is given by:*

$$oe_{r^a_{bin}}(P) = \begin{cases} \frac{p_P \cdot n_\Delta}{p_P + n_\Delta}, & \text{if } a = 1 \\ \frac{n_\Delta}{1 + n_\Delta}, & \text{if } a = 0 \\ \frac{\hat{p}^a \cdot n_\Delta}{\hat{p} + n_\Delta}, \text{ with } \hat{p} = \min(\frac{a \cdot n_\Delta}{1-a}, p_P), & \text{else} \end{cases}$$

Proof. Let S be any specialization of P and $G = \gamma(S)$ the generalization with $\Delta(G,S) \subseteq \Delta(P',P'')$, which exists according to the previous lemma, since S is also a specialization of P'. The number of negatives in G is equal to the number of negatives covered by S plus the number of negatives, which are covered by G, but not by S: $n_G = n_S + n_{\Delta(G,S)}$. By construction it holds that $n_{\Delta(G,S)} \leq n_\Delta$. Additionally, we can assume $p_S > 0$, that is, S contains at least one positive instance, since $r^a_{bin}(S) \leq 0$ otherwise.

In the proof, we will first derive an upper bound that depends on the number of positives in the specialization S, which is unknown at the time P is evaluated. In a second step we therefore determine the maximum value of this function. The quality of S is given by:

$$
\begin{aligned}
r^a_{bin}(S) &= (p_S + n_S)^a \cdot (\tau_S - \max_{H \subset S} \tau_H) && (1)\\
&\leq (p_S + n_S)^a \cdot (\tau_S - \tau_G) && (2)\\
&= (p_S + n_S)^a \cdot (\frac{p_S}{p_S + n_S} - \frac{p_G}{p_G + n_S + n_{\Delta(G,S)}}) && (3)\\
&\leq (p_S + n_S)^a \cdot (\frac{p_S}{p_S + n_S} - \frac{p_S}{p_S + n_S + n_{\Delta(G,S)}}) && (4)\\
&= (p_S + n_S)^a \cdot (\frac{p_S \cdot (p_S + n_S + n_{\Delta(G,S)}) - (p_S \cdot (p_S + n_S))}{(p_S + n_S)(p_S + n_S + n_{\Delta(G,S)})}) && (5)\\
&= \frac{p_S \cdot n_{\Delta(G,S)}}{(p_S + n_S)^{1-a}(p_S + n_S + n_{\Delta(G,S)})} && (6)\\
&\leq \frac{p_S \cdot n_{\Delta(G,S)}}{(p_S)^{1-a}(p_S + n_{\Delta(G,S)})} && (7)\\
&= \frac{p^a_S \cdot n_{\Delta(G,S)}}{(p_S + n_{\Delta(G,S)})} && (8)\\
&\leq \frac{p^a_S \cdot n_\Delta}{(p_S + n_\Delta)} := f^a(p_S) && (9)
\end{aligned}
$$

The transformation to line 2 is possible, since $G \subset S$. In line 4 it is used that $p_S \leq p_G$, as the positives of S are a subset of the positive of its generalization G. In line 7 it is exploited that the denominator is strictly increasing with increasing n_S, because $1 - a \in [0;1]$. Therefore, the smallest denominator and thus the largest value for the overall term is achieved by setting $n_S = 0$. The term in line 8 is strictly increasing as a function of $n_{\Delta(G,S)}$. Since $n_{\Delta(G,S)} \leq n_\Delta$, line 9 follows.

In the final line 9, the function $f^a(p_S)$ is defined, which provides an upper bound on the interestingness of P that depends on the number of positives within the specialization. This number is not known, when the pattern P is evaluated. Intuitively, for large number of positives in the specialization removing n_Δ negative instances will not change the target share in the subgroup much, therefore the interestingness of the generalization is limited. On the other hand, for small numbers of positive instances S is overall small and possibly not interesting for that reason. p_S is at least 1, since S otherwise is not interesting anyway and at most p_P, as the number of positives for S is smaller than for its generaliza-

tion P. Next, we analyze for which value of p_S the function $f^a(p_S)$ of line 9 reaches its maximum in the interval $[1; p_P]$. This depends on the parameter a of the interestingness measure:

1. For $a = 1$ it holds that $f^1(p_S) = \frac{p_S \cdot n_\Delta}{p_S + n_\Delta}$. This function is strictly increasing in p_S. That is, the more positive instances are contained in S, the higher is the derived upper bound. The maximum is reached at highest value in the domain of definition: $\max(f^1(p_S)) = f^1(p_P) = \frac{p_P \cdot n_\Delta}{p_P + n_\Delta}$.
2. In contrast for $a = 0$, $f^0(p_S) = \frac{n_\Delta}{p_S + n_\Delta}$ is strictly decreasing. Thus, the maximum value of f^0 is reached for $p_S = 1$, the minimum possible value of p_S: $\max(f^0(p_S)) = f^0(1) = \frac{n_\Delta}{1 + n_\Delta}$.
3. For $0 < a < 1$, f^a reaches a maximum for a certain value p^* within the domain of definition. To determine that, we compute the first derivative of f^a using the quotient rule.

$$\begin{aligned}\frac{d}{d\,p_S} f^a(p_S) &= n_\Delta \cdot \frac{d}{dp_S} \frac{p_S^a}{p_S + n_\Delta} \\ &= n_\Delta \frac{(n_\Delta + p_S) \cdot a \cdot p_S^{a-1} - p_S^a}{(n_\Delta + p_S)^2} \\ &= n_\Delta \cdot p_S^{a-1} \frac{a n_\Delta + a \cdot p_S - p_S}{(n_\Delta + p_S)^2} := (f^a)'\end{aligned}$$

   The only root of this derivative is at $p^* := \frac{a \cdot n_\Delta}{1 - a}$. As can be easily shown, $(f^a)'(p_S)$ is greater than zero for p_S smaller than p^* and lower than zero for p_S greater than p^*. Therefore, p^* is the only maximum of $f^a(p_S)$. Thus, if $p_P > p^*$, then p^* is the maximum value of f^a, otherwise the maximum is reached at the highest value of the domain of definition: $\max(f^a(p_S)) = f^a(\hat{p}) = \frac{\hat{p}^a \cdot n_\Delta}{\hat{p} + n_\Delta}$, with $\hat{p} = \min(\frac{a \cdot n_\Delta}{1 - a}, p_P)$.

   Overall, for any specialization S it holds that $r^a_{bin}(S) \leq f^a(p_S) \leq \max f^a(p_S) = oe_{r^{bin}_a}(P)$, with the function maxima as described above, therefore $oe_{r^{bin}_a}(P)$ as defined in the theorem is a correct optimistic estimate. □

For any pair of generalizations of P (P' and P'') as well as for any pair of P ($P' = P$) and one of its generalization (P''), this theorem provides an optimistic estimate of P. The optimistic estimate bound is dependent on the number of positives in the subgroup and the difference of negative instances between P' and P''. It is low, if either there are only few positives in P or the difference of negative instances between the pair of generalizations is small (or a combination of both). Since the number of positives in P is independent of the chosen pair P', P'', the pair with the minimum difference of negative instances implies the tightest upper bound, which should be used to maximize the effects of pruning.

As a special case the theorem includes that the interestingness of any pattern is ≤ 0, if n_Δ is 0. To the authors knowledge, it is the first measure that includes these differences in optimistic estimate bounds for subgroup discovery.

4.4 Difference-based Optimistic Estimates for Numeric Targets

Next, we will show that a related approach can be used to obtain optimistic estimates for generalization-aware interestingness $r^a_{num} = i_P^a \cdot (\mu_P - \max_{H \subset P} \mu_H)$ in settings with numeric target concepts.

Theorem 3. *In a task with a numeric target concept, consider the pattern P with i_P instances and a maximum target value of $\max_P$. $P' \subseteq P$ is either P itself or one of its generalizations and $P'' \subset P'$ is a generalization of P'. Let $i_\Delta = |\Delta(P'', P')|$ be the number of instances contained in P'', but not in P' and $\min_\Delta$ the minimum target value contained in $\Delta(P'', P')$. Then, an optimistic estimate of P for the generalization aware quality function r^a_{num} is given by:*

$$oe_{r^a_{num}}(P) = \max(0, oe'_{r^a_{num}}(P)),$$

$$oe_{r^a_{num}}(P)' = \begin{cases} \frac{i_\Delta \cdot i_P}{i_P + i_\Delta} \cdot (\max_P - \min_\Delta), & \text{if } a = 1 \\ \frac{i_\Delta}{1 + i_\Delta} \cdot (\max_P - \min_\Delta), & \text{if } a = 0 \\ \frac{\hat{i}^a \cdot i_\Delta}{\hat{i} + i_\Delta} \cdot (\max_P - \min_\Delta), \text{ with } \hat{i} = \min(\frac{a \cdot i_\Delta}{1-a}, i_P), & \text{else} \end{cases}$$

Proof. We consider any specialization $S \supset P$ and its generalization $G = \gamma(S)$ according to Lemma 1. Then we can estimate the interestingness of S:

$$r^a_{num}(S) = {i_S}^a \cdot (\mu_S - \max_{H \subset S} \mu_H) \tag{1}$$

$$\leq {i_S}^a \cdot (\mu_S - \mu_G) \tag{2}$$

$$= {i_S}^a \cdot \left(\frac{\sum_{i \in sg(S)} tc(i)}{i_S} - \frac{\sum_{i \in sg(S)} tc(i) + \sum_{j \in \Delta(G,S)} tc(j)}{i_S + i_{\Delta(G,S)}}\right) \tag{3}$$

$$= {i_S}^{a-1} \cdot \left(\sum_{i \in sg(S)} tc(i) - \frac{i_S \cdot (\sum_{i \in sg(S)} tc(i) + \sum_{j \in \Delta(G,S)} tc(j))}{i_S + i_{\Delta(G,S)}}\right) \tag{4}$$

$$= {i_S}^{a-1} \cdot \left(\frac{i_{\Delta(G,S)} \sum_{i \in sg(S)} tc(i) - i_S \sum_{j \in \Delta(G,S)} tc(j)}{i_S + i_{\Delta(G,S)}}\right) \tag{5}$$

$$\leq {i_S}^{a-1} \cdot \left(\frac{i_{\Delta(G,S)} \cdot i_S \cdot \max_{i \in S} tc(i) - i_S \cdot i_{\Delta(G,S)} \cdot \min_{j \in \Delta(G,S)} tc(j)}{i_S + i_{\Delta(G,S)}}\right) \tag{6}$$

$$= \frac{i_{\Delta(G,S)} \cdot {i_S}^a}{i_S + i_{\Delta(G,S)}} \cdot (\max_{i \in S} tc(i) - \min_{j \in \Delta(G,S)} tc(j)) \tag{7}$$

$$\leq \frac{i_\Delta \cdot {i_S}^a}{i_S + i_\Delta} \cdot (\max_P - \min_\Delta) = f(i_S) \cdot (\max_P - \min_\Delta) \tag{8}$$

In line 2 it is used that G is a generalization of S, then it is exploited that $sg(G) = sg(S) \cup \Delta(G,S), S \cap \Delta(G,S) = \emptyset$. In line 6 we utilize that the sum of any set of values is bigger than the minimum appearing value times the size of the set, but smaller than the maximum appearing value times the size of the set. Line 8 uses that $i_{\Delta(G,S)} \leq i_\Delta$.

f^a is a function over the unknown number of all instances in the specialization, which can be any number in $[1; i_P]$. f^a is always positive. Therefore, if $(\max_P - \min_\Delta) \leq 0$, the optimistic estimate is given by 0. Else, the maxima of f^a, which have already been derived in the proof of Theorem 2, determine the bound: $f^a(i_S)$ is strictly increasing for $a = 1$, strictly decreasing for $a = 0$ and reaches a maximum at $\frac{a \cdot i_\Delta}{1-a}$ or at i_P otherwise. Thus: $r^a_{num}(S) \leq (f^a(i_S)) \cdot (\max_P - \min_\Delta) \leq \max(f^a(i_S) \cdot (\max_P - \min_\Delta)$. The bounds follow directly from the inserting the resp. maxima values. Since this holds for any specialization S of P, $oe_{r^a_{num}}(P)$ is a correct optimistic estimate for P. □

Similar to the optimistic estimate in the binary case, the derived optimistic estimate is low, if either the number of instances covered by P is low, or if the difference in the number of instances covered between the generalizations P'' and P' is low (or a combination of both). However additionally, the bound also considers the range of the target variable in these patterns, that is, the maximum occurring target value in P and the minimum target value in the difference set of instances. As a result, the bound gets zero, if the minimum target value removed by adding a selector to a generalization of P was higher than the maximum remaining target value in P.

5 Algorithm

The presented optimistic estimates can in general be applied in combination with any search strategy. In this paper we focus on adapting an exhaustive algorithm, i.e., apriori [1,16]. This approach is especially suited for the task of generalization-aware subgroup discovery, since its levelwise search strategy guarantees that specializations are always evaluated after their generalizations and the highest target share found in generalizations can efficiently be propagated from generalizations to specializations, see [5]. Therefore, and for better comparability with previous approaches, we chose apriori as a basis for our novel algorithm. Using the following adaptations the algorithm is not only capable of determining the proposed optimistic estimates. The algorithm also propagates the required information very efficiently. Due to limited space, we will not describe the base algorithm, which has been extensively described in literature [1,20,16,5], but instead focus only on the differences. We start by describing the binary case.

Apriori performs a levelwise search, where new candidate patterns are generated from the last level of more general patterns. In our adaptation of the algorithm additional information is stored for each candidate. This includes the maximum target share in generalizations of this pattern, the minimum number of negatives covered by any generalization and the minimum number of negatives that were removed in generalizations of this pattern. After the evaluation of a pattern the number of positives, the number of negatives and the resulting target share are additionally saved in each candidate. The minimum number of negative instance in a generalization is required to compute the minimum

number of instances, which are contained in the pattern, but not in a generalization. The other statistics are directly required to compute either the quality or the optimistic estimates of the pattern. Whenever a new candidate pattern P is generated in apriori, it is checked for all its direct generalizations G, if it is contained in the last levels candidate set. During this check, the statistics for the maximum target share in generalizations, the minimum number of negatives in a generalization and the minimum number of negatives that were removed in any generalization of this pattern can be computed by using the information stored in the generalizations and simple minimum/maximum functions. In doing so, the statistics required to compute the quality of the pattern and the optimistic estimates are propagated very efficiently from one level of patterns to the next level of more specific patterns.

In the evaluation phase (the counting phase in classical apriori) each candidate is evaluated. This requires to determine the coverage of the pattern. Combined with previously computed statistics about generalizations this is used to compute the interestingness according to the chosen generalization-aware measure. Subgroups with sufficient high score are placed in the result set, potentially replacing others in a top-k approach. Afterwards the target share in generalizations and the minimum number of removed negative instances are updated by using the statistics of the current patterns coverage. After the evaluation of a pattern all optimistic estimates, that is, traditional estimates (see theorem 1) and difference-based estimates are computed from the information stored for a candidate. If any optimistic estimate is lower than the threshold given by the result set for a top-k pattern, then the pattern is removed from the list of current candidates. Thus, no specializations of this pattern are explored in the next level of search.

The approach for numeric target concepts is very similar, except that minimum/maximum and mean target values as well as overall instance counts of the candidate patterns are stored instead of counts of positives and negatives. When determining the pruning bounds, a pattern is compared with all its direct generalizations. For each generalization an optimistic estimate bound is computed based on the difference of instances between the generalization and the specialization and the stored minimum/maximum target values. The tightest bound can be applied for pruning.

For the experiments, the algorithm was implemented in the open-source environment VIKAMINE [3]. The implementation utilizes an efficient bitset-based data structure to determine the coverage of patterns efficiently.

6 Evaluation

In this section, we show the effectiveness of the presented approach in experiments using well-known datasets from the UCI [9] repository. As a baseline algorithm we use a variant of the MPR-algorithm presented in [5], as this is the most recently proposed algorithm for this task. The algorithm was slightly modified to support top-k mining and to incorporate the bounds of Theorem 1

for any a. Since this algorithm follows the same search strategy as our novel algorithm, that is, apriori, it allows to determine the improvements that originate directly from the advanced pruning bounds presented in this paper. Results below are shown for $k = 20$, a realistic number for practical applications, which was also used for example as beam size in [26]. Different choices of k lead to similar results. For the numeric attributes an equal-frequency discretization was used, using all half-open intervals from the cutpoints as selectors. The experiments were performed on an office PC with 2.8 Ghz and 6 GB RAM.

In the first part of the evaluation we investigated the setting of a binary target concept using different generalization-aware quality functions r^a_{bin}. We compared the runtimes of the presented algorithm with traditional pruning only and with the novel generalization-aware bounds. The results show, that utilizing difference-based pruning leads to significant runtime improvements in almost all tasks, see Table 1. The improvements range from a factor of about 2 to over 20 in the datasets hypothyroid, audiology and spammer. For a more detailed analysis we investigated these tasks more closely. It turned out that the search space for these datasets contained multiple selectors that covered a vast majority of the instances. Conjunctive combinations of subsets of these selectors still cover a large part of the dataset and especially of the positive instances. As traditional optimistic estimates are based on this number of covered positive instance, pruning cannot be applied on these combinations efficiently. In contrast, since the number of negative instances, by which those patterns differ from generalizations, is often very low in these cases, such combination can be pruned often using the difference-based optimistic estimates presented in this paper. This leads to the massive improvements. We can conclude that our new pruning scheme is especially efficient, if many selectors cover a majority of the dataset. In some cases the algorithms did not finish due to out of memory errors despite the large amount of available memory. This does occur less often using the novel bounds, see for example the results for the vehicle dataset, since less candidates are generated in apriori, if more advanced bounds are applied.

In the second part of the evaluation the interestingness measure $r^{0.5}_{bin}$, a generalization-aware variant of the binomial-test, was further analyzed by comparing the runtimes for different search depth (maximum number of selectors in a pattern), see Table 2. As before, almost all tasks finished earlier using the novel difference-based pruning. While the improvement is only moderate for low search depth, massive speedups can be observed for $d = 5$ and $d = 6$. For $d = 6$ many algorithms with only traditional pruning did not finish because of limited memory. When additionally using the novel bounds, this happened only in two datasets, as less candidates were generated.

In the last part of the evaluation the improvements in a setting with numeric target concepts and quality functions q^a_{num} were examined. For subgroup discovery with numeric targets and generalization-aware quality functions no optimistic estimates have been proposed so far. To allow for a comparison nonetheless, we use the optimistic estimate bound $\bar{oe}^1_{num} = \sum_{x:tc(x)>\mu_\emptyset}(tc(x) - \mu_\emptyset)$, which has been shown to be a correct optimistic estimate for q^1_{num}. Since $r^a_{num}(P) \leq$

Table 1. Runtime comparison (in s) of the base algorithm with traditional pruning based on the positives (std) and the novel algorithm with additional difference-based pruning (dbp) using different size parameters a for quality functions r^a_{bin}. The maximum number describing selectors was limited to $d = 5$. "-" indicates that the algorithm did not finish due to lack of memory.

a	0.0		0.1		0.5		1.0	
pruning	dpb	std	dpb	std	dpb	std	dpb	std
adults	1.1	1.0	17.8	48.7	1.6	8.1	1.0	1.7
audiology	0.2	62.3	24.9	51.6	0.6	51.7	0.1	57.4
census-kdd	16.4	16.2	-	-	107.9	2954.3	18.5	94.0
colic	<0.1	<0.1	1.7	4.8	0.4	5.1	0.1	1.2
credit-a	<0.1	<0.1	2.6	4.1	1.2	3.6	<0.1	0.4
credit-g	0.2	0.2	24.4	42.5	4.0	35.2	0.4	4.6
diabetes	1.0	3.8	5.9	12.6	1.2	9.3	<0.1	0.7
hepatitis	1.5	11.8	2.3	4.9	0.8	3.3	<0.1	0.5
hypothyroid	0.1	1.2	2.0	37.1	1.7	39.0	<0.1	21.2
spammer	4.3	5.5	133.0	-	29.3	172.2	0.5	27.6
vehicle	2.3	2.7	-	-	15.6	-	0.9	-

Table 2. Runtime comparison (in s) of the base algorithm with traditional pruning based on the positives (std) and the novel algorithm with additional difference-based pruning (dbp) using different maximum numbers d of describing selectors in a pattern. As quality functions the generalization-aware mean test $r^{0.5}_{bin}$ was used. "-" indicates that the algorithm did not finish due to lack of memory.

d	3		4		5		6	
pruning	dpb	std	dpb	std	dpb	std	dpb	std
adults	1.0	1.1	0.9	1.8	1.6	8.1	1.7	30.2
audiology	0.1	0.1	0.1	2.8	0.6	51.7	-	-
census-kdd	17.9	20.6	37.2	99.8	107.9	2954.3	267.5	-
colic	0.1	0.2	0.3	1.1	0.4	5.1	0.4	16.4
credit-a	0.1	0.1	0.3	0.7	1.2	3.6	1.2	12.9
credit-g	0.2	0.2	1.5	4.0	4.0	35.2	7.0	-
diabetes	0.1	0.1	0.5	1.3	1.2	9.3	2.0	67.1
hepatitis	<0.1	0.1	0.2	0.6	0.8	3.3	0.3	11.9
hypothyroid	0.1	0.2	0.5	2.7	1.7	39.0	-	-
spammer	1.3	1.6	5.7	15.5	29.3	172.2	88.3	-
vehicle	1.0	1.3	4.8	57.8	15.6	-	-	-

$r^1_{num}(P) \leq q^1_{num}(P)$ this can also be used as a (non-tight) optimistic estimate for any generalization-aware quality function r^a_{num}. Results are shown in Table 3. Since the applied traditional bound is tight for $a = 1$, the runtimes in this case are relatively low already for the studied datasets, leaving only little room for improvement. For lower values of a, significant runtime improvements can be observed, which reach a full order of magnitude (e.g., for the datasets concrete_data and housing). The relative runtime improvement is on average highest for $a = 0.5$. This can be explained by the fact that for lower values of a even

small subgroups can be considered as interesting. This makes it more difficult to exclude subgroups by pruning also when using the difference-based bounds.

Table 3. Runtime comparison (in s) of the base algorithm with traditional pruning based on the positives (std) and the novel algorithm with additional difference-based pruning (dbp) using different size parameters a for quality functions r^a_{num} for numeric target concepts. The maximum number describing selectors was limited to $d = 5$.

a	0.0		0.1		0.5		1.0	
pruning	dpb	std	dpb	std	dpb	std	dpb	std
adults	19.6	92.5	22.5	89.6	14.8	64.7	3.9	14.9
concrete_data	4.7	20.8	6.2	20.1	1.3	11.2	0.1	0.3
credit-a	3.7	14.1	5.1	13.8	3.1	9.7	0.4	0.8
credit-g	6.6	53.0	8.5	54.5	7.9	40.3	0.5	1.0
diabetes	5.6	20.4	8.5	18.6	5.2	15.0	0.3	0.7
forestfires	2.3	10.4	3.4	11.1	2.7	9.6	2.4	6.5
heart-c	3.5	17.5	5.6	17.5	2.9	13.2	0.2	0.5
housing	2.0	28.2	3.4	26.4	1.7	23.8	0.1	3.0
yeast	3.1	14.3	3.5	13.9	1.5	8.4	0.1	0.8

7 Conclusions

In this paper we proposed a new scheme of deriving optimistic estimates bounds for subgroup discovery with interesting measures that take statistics of generalizations into account. In contrast to previous approaches the bounds are not only based on the anti-monotonicity of instances, which are contained within the subgroup, but also on the number of instance that are covered by a pattern, but not by its generalization. The optimistic estimates have been incorporated in an efficient algorithm that outperforms previous approaches by up to an order of magnitude. The speed-up is especially high, if the dataset contains selection expressions that cover a large part of the dataset.

In the future we plan to extend this approach to explore novel interestingness measures that take generalizations into account. Furthermore, an analysis of different search strategies, e.g., reverse-depth-first search, for this task is an interesting direction.

References

1. Agrawal, R., Imielienski, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record (May), 1–10 (1993)
2. Atzmueller, M., Lemmerich, F.: Fast subgroup discovery for continuous target concepts. Foundations of Intelligent Systems (2009)
3. Atzmueller, M., Lemmerich, F.: VIKAMINE–Open-Source Subgroup Discovery, Pattern Mining, and Analytics. Proceedings of the European conference on machine learning and knowledge discovery in databases pp. 4–7 (2012)
4. Aumann, Y., Lindell, Y.: A statistical theory for quantitative association rules. Knowledge Discovery and Data Mining pp. 261–270 (1999)

5. Batal, I., Hauskrecht, M.: A concise representation of association rules using minimal predictive rules. Machine Learning and Knowledge Disc. pp. 87–102 (2010)
6. Batal, I., Hauskrecht, M.: Constructing classification features using minimal predictive patterns. Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing pp. 869–877 (2010)
7. Bay, S., Pazzani, M.: Detecting change in categorical data: Mining contrast sets. Proceedings of the fifth ACM SIGKDD int. conf. on KDD (1999)
8. Bayardo, R.: Efficiently mining long patterns from databases. ACM Sigmod Record pp. 85–93 (1998)
9. Blake, C., Merz, C.J.: {UCI} Repository of machine learning databases (1998)
10. Cheng, H., Yan, X., Han, J., Yu, P.: Direct discriminative pattern mining for effective classification. ICDE '08 Proceedings of the 2008 IEEE 24th International Conference on Data Engineering pp. 169–178 (Apr 2008)
11. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining pp. 1–11 (1999)
12. Garriga, G., Kralj, P., Lavrac, N.: Closed sets for labeled data. The Journal of Machine Learning Research 9, 559–580 (2008)
13. Geng, L., Hamilton, H.J.: Interestingness measures for data mining. ACM Computing Surveys 38(3), 9–es (Sep 2006)
14. Grosskreutz, H., Boley, M., Krause-Traudes, M.: Subgroup discovery for election analysis: a case study in descriptive data mining. Disc. Science pp. 57–71 (2010)
15. Grosskreutz, H., Rüping, S., Wrobel, S.: Tight optimistic estimates for fast subgroup discovery. Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases pp. 440–456 (2008)
16. Kavšek, B., Lavrač, N.: Apriori-Sd: Adapting Association Rule Learning To Subgroup Discovery, vol. 20 (Sep 2006)
17. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining. pp. 249–271. American Association for Artificial Intelligence (1996)
18. Kralj Novak, P., Lavrač, N., Webb, G.I.: Supervised Descriptive Rule Discovery : A Unifying Survey of Contrast Set , Emerging Pattern and Subgroup Mining 10, 377–403 (2009)
19. Lemmerich, F., Puppe, F.: Local Models for Expectation-Driven Subgroup Discovery. 2011 IEEE 11th International Conference on Data Mining pp. 360–369 (2011)
20. Morishita, S., Sese, J.: Traversing Itemset Lattices with Statistical Metric Pruning. In Proc. of ACM SIGMOD pp. 226–236 (2000)
21. Nijssen, S., Guns, T., Raedt, L.D.: Correlated itemset mining in roc space: a constraint programming approach. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (2009)
22. Webb, G.I.: OPUS: An efficient admissible algorithm for unordered search. arXiv preprint cs/9512101 3, 431–465 (1995)
23. Webb, G.I.: Discovering associations with numeric variables. Proceedings of the seventh ACM SIGKDD int. conf. on Knowledge discovery and data mining (2001)
24. Webb, G.I., Zhang, S.: Removing trivial associations in association rule discovery. Proceedings of the First International NAISO Congress on Autonomous Intelligent Systems, p. NAISO Academic Press: Geelong, 2002 (2002)
25. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. Principles of Data Mining and Knowledge Discovery (1997)
26. Zimmermann, A., Raedt, L.D.: From Subgroup Discovery to Clustering. Machine Learning 77(1), 125—-159 (2009)