Difference-based Estimates for Generalization-aware Subgroup Discovery

Extended abstract of a paper originally published on ECML/PKDD 2013

Florian Lemmerich, Martin Becker, and Frank Puppe
University of Würzburg
{lemmerich, becker, puppe}@informatik.uni-wuerzburg.de

Abstract

In this work, we approach the topic of efficient subgroup mining with interestingness measures, which also take statistics on generalizations of the subgroup into account. For this setting we develop new optimistic estimate bounds, which allow to safely prune large parts of the search space. In contrast to previous approaches, the bounds are not only based on the anti-monotonicity of the number of covered instances of a pattern, but also on the number of instances, by which a pattern differs in comparison to its generalizations. Incorporating these bounds in an efficient algorithms leads to runtime improvements of up to an order of magnitude.

1 Problem Setting

Subgroup discovery [5] is a key technique for descriptive data mining. It aims at identifying descriptions of subsets of the data that show an interesting behavior with respect to a certain target concept. This is accomplished by using an interestingness measure to assign a score to all candidate patterns in the search space of all conjunctive descriptions. Traditional measures are based on the statistics of the evaluated subgroup and the entire dataset. In particular, the most popular family of interestingness measures weights between the number of instances covered by the subgroup, and the difference of the target share (or target mean value in a numeric target setting) in the subgroup a the target share in the total population. In recent research [1; 2; 3] these measure have been adapted to obtain more interesting and less redundant results: Generalization-aware measures replace the comparison with the target share (mean value) in the overall dataset with a comparison to the maximum target share of all generalizations of the subgroup. E.g., to compute the interestingness score of the subgroup $A \wedge B$, the target share for the three subgroup patterns $\emptyset$, A, B are compared to the target share of $A \wedge B$. In this paper, we focus on the most important families of interestingness measures for nominal and numeric target concepts in this direction:

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

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

Here, i_P is the number of instances covered by the subgroup P, $\tau_P(\mu_P)$ is the target share (target mean value) in the subgroup P and $\max_{H \subset P} \tau_H(\max_{H \subset P} \mu_H)$ is the maximum target share (target mean value) in all generalizations of P. a is an user-specified parameter that allows to weight between the two factors.

This paper does not argue about the usefulness of these adaptations, but focuses on efficient subgroup mining in this setting. In particular, we propose novel, tighter *optimistic estimate bounds* [5] that allow to prune parts of the search space without losing the optimality of the results: The basic idea of optimistic estimates the following: if one can guarantee that no specialization of the currently evaluated pattern will have an interestingness score which is good enough to include the respective pattern into the result set then we can safely omit these patterns from the search. In this regard, we aim at the strictest bounds possible to reduce the remaining search space and thus to speed up the search process.

2 Difference-based estimates

Previous approaches to compute optimstic estimates are almost exclusively based on the anti-monotonicity of covered (positive) instances: For instance, if the subgroup A covers 10 positive examples, then each specialization of A, e.g., $A \wedge X$ covers also at most 10 positive examples. In addition to the statistics of the currently evaluated subgroup, our approach also takes into account statistics of generalizations in order to to determine the interestingness score. This additional information is used to determine tighter optimistic estimates.

For this end, the following lemma is proposed:

Lemma. *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 $\Delta(\gamma(S), S) \subseteq \Delta(A, B)$.*

The lemma exploits, what can be described as an *anti-monotonicity of differences* in comparison to generalizations. For example, assume there are 5 instances, which are covered by U, but not by $U \wedge V$. Then the lemma guarantees, that for each specialization $S = U \wedge V \wedge X \wedge \ldots \wedge Y$ there exists a generalization, such that the difference between this generalization and S is also at most 5 instances (cf. also [4]).

Now, consider the interestingness score of such a specialization S: If S covers only few instances, then by the definition of the used interestingness measures, S is of low interestingness. On the other hand, if S covers more instances, the increase of the target share

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	-	-	-

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 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.

in comparison to its generalization $\gamma(S)$ is limited by the lemma, since it states that $\gamma(S)$ only covers at most 5 more negative instances than S. As a consequence S is also not interesting in this case.

These considerations are exploited in formal theorems, which allow to determine optimistic estimates based on the difference of instances in generalizations:

Theorem. *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:*

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

This theorem provides optimistic estimate bounds, which are tight (low), if either (1) the number positives covered by a subgroup is low, or (2) if the difference of negatives between the subgroup and a generalization is low, or (3) if the difference of negatives between a generalization of the subgroup and another generalization of this generalization is low.

Another theorem (not shown in this abstract) introduces optimistic estimate bounds for the setting with a numeric target setting and mean-based generalization-aware interestingness measures r^a_{num}. These bounds also exploit the difference of the minimum target value removed in a specialization step to the maximum target value remaining in the subgroup.

3 Algorithm

Although the proposed optimistic estimate bounds can in principal be applied with any search strategy, we focus in this work on adapting Apriori, which is also employed by the current state-of-the-art algorithm of this problem setting [1]. For each candidate pattern additional information is stored, e.g., the minimum number of negatives in a generalization, the minimum difference in coverage between two generalizations and the maximum target share in a generalization of this pattern. The information is propagated efficiently during candidate generation and updated during the evaluation of the subgroup.

4 Evaluations

The effectiveness of the difference-based optimistic estimate bounds and its incorporation in an algorithm was evaluated in several series of experiments. Exemplary results are shown in Table 1. It can be observed that the novel approach improves the runtime often of more than an order of magnitude. Further investigation showed that the runtime improvement is particularly large, if the dataset contains many selectors that cover large parts of the dataset (see e.g., the audiology dataset). In can also be seen, that out-of-memory errors occur less often using the improved bounds, since less candidates are generated.

The full paper includes formal proofs, a more detailed algorithm description and more experimental results. It has been published as: Florian Lemmerich, Martin Becker, Frank Puppe: Difference-Based Estimates for Generalization-Aware Subgroup Discovery. In: Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, Filip Zelezný (Eds.): Proceedings of ECML/PKDD 2013, Part III, pages 288-303.

References

[1] Batal, I., Hauskrecht, M.: A concise representation of association rules using minimal predictive rules. Machine Learning and Knowledge Disc. pp. 87–102 (2010)

[2] 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)

[3] Lemmerich, F., Puppe, F.: Local Models for Expectation-Driven Subgroup Discovery. 2011 IEEE 11th International Conference on Data Mining pp. 360–369 (2011)

[4] 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)

[5] Wrobel, S.: An algorithm for multi-relational discovery of subgroups. Principles of Data Mining and Knowledge Discovery (1997)