DMKD manuscript No.
(will be inserted by the editor)

Fast Exhaustive Subgroup Discovery with Numerical Target Concepts

Florian Lemmerich · Martin Atzmueller · Frank Puppe

Received: date / Accepted: date

Abstract Subgroup discovery is a key data mining method that aims at identifying descriptions of subsets of the data that show an interesting distribution with respect to a pre-defined target concept. For practical applications the integration of numerical data is crucial. Therefore, a wide variety of interestingness measures has been proposed in literature that use a numerical attribute as the target concept. However, efficient mining in this setting is still an open issue. In this paper, we present novel techniques for fast exhaustive subgroup discovery with a numerical target concept. We initially survey previously proposed measures in this setting. Then, we explore options for pruning the search space using optimistic estimate bounds. Specifically, we introduce novel bounds in closed form and *ordering-based bounds* as a new technique to derive estimates for several types of interestingness measures with no previously known bounds. In addition, we investigate efficient data structures, namely adapted FP-trees and bitset-based data representations, and discuss their interdependencies to interestingness measures and pruning schemes. The presented techniques are incorporated into two novel algorithms. Finally, the benefits of the proposed pruning bounds and algorithms are assessed and compared in an extensive experimental evaluation on 24 publicly available datasets. The novel algorithms reduce runtimes consistently by more than one order of magnitude.

Florian Lemmerich
GESIS – Leibniz Institute for the Social Sciences
Computational Social Science Department, Cologne, Germany
E-mail: florian.lemmerich@gesis.org

Martin Atzmueller
University of Kassel, Research Center for Information System Design (ITeG),
Knowledge and Data Engineering Group, Germany
E-mail: atzmueller@cs.uni-kassel.de

Frank Puppe
University of Würzburg, Institute of Computer Science,
Artificial Intelligence and Applied Computer Science Group, Germany
E-mail: puppe@informatik.uni-wuerzburg.de

Preprint of: Florian Lemmerich, Martin Atzmueller, and Frank Puppe (2015): Fast Exhaustive Subgroup Discovery with Numerical Target Concepts. Data Mining and Knowledge Discovery

Keywords subgroup discovery · pattern mining · numerical data · pruning · data structures · data mining · algorithms

1 Introduction

Subgroup discovery aims at identifying descriptions of subsets of the data that deviate from the overall dataset with respect to a certain property of interest, often also called target concept. As an established method of data mining, it has been well-examined concerning binary target concepts with a finite number of possible values, see for example (Klösgen 1996; Wrobel 1997; Lavrač et al. 2004). However, practical applications often involve numerical data, i.e., attributes with a continuous domain. In this context, this paper focuses specifically on the setting in which the target concept is given by a numerical attribute. Transforming this problem setting to the standard binary one by using discretization techniques (Fayyad and Irani 1993; Kotsiantis and Kanellopoulos 2006) in a pre-processing step can lead to a (possibly crucial) loss of information, cf. (Moreland and Truemper 2009; Freidlin and Gastwirth 2000). Therefore, a broad variety of interestingness measures which directly consider the distribution of a numerical target attribute has been proposed for pattern evaluation in the literature. Efficient subgroup discovery using these measures is an open issue since the transfer of techniques developed for the binary target setting is challenging. In this paper, we discuss in particular methods for exhaustive mining with guaranteed optimal results.

Beside the search strategy, discovery algorithms are characterized by the used *data structures* and the applied *pruning schemes* that allow for skipping parts of the search space in the discovery algorithm. An essential pruning technique that guarantees the optimality of results is *optimistic estimate pruning*. It substantially reduces the number of required subgroup evaluations based on the following principle: if it can be proven, that no specialization of the currently investigated subgroup is interesting enough (according to the chosen interestingness measure) to be included in the result set of subgroups, then we can skip the evaluation of all these specializations.

In this paper, we first survey previously proposed interestingness measures for numerical properties of interest from literature, including mean-based, variance-based, median-based, and rank-based interestingness measures as well as a measure based on the Kolmogorov-Smirnov statistical test. For these interestingness measures, we present a large set of optimistic estimate bounds that can be used for pruning the search space. In that direction, we propose the formalism of interestingness measures that are *estimable by ordering* of the target values as a means to derive optimistic estimates for a variety of interestingness measures with no previously known bounds. For faster computation, we additionally introduce approximations that can be computed knowing only

This paper summarizes and extends contents of the dissertation of the first author (Lemmerich 2014). A small part of this work, i.e., the *SD-Map** algorithm for mean-based interestingness measures only, was previously described in (Atzmueller and Lemmerich 2009).

a subset of the subgroup instances. Since ordering-based bounds cannot be determined with all data structures, we also discuss bounds in closed form that require only few subgroup statistics. Besides optimistic estimate pruning, we show how popular data structures for subgroup discovery, i.e., FP-trees and bitset-based vertical data structures, can be transfered from binary to numerical target concepts. The proposed techniques are incorporated into two practical algorithms. An extensive experimental evaluation of the discussed bounds and algorithms on 24 publicly available datasets show substantial runtime improvements.

This paper does not consider the overall subgroup discovery process, but concentrates only on exact solutions of the central algorithmic step, ignoring heuristic approaches for this task. Additionally, we assume that the set of describing selectors for the search has already been determined beforehand. However, we acknowledge that this can be a challenging task that also possibly involves loss of information especially considering numerical attributes. Furthermore, we focus exclusively on classic interestingness measures that are based only on the statistics of the evaluated subgroups and do not take the subgroup description into account.

The rest of the paper is structured as follows: Section 2 discusses related work. Next, Section 3 describes the basics of subgroup discovery with an emphasis on numerical target concepts and introduces the used notations. Then, Section 4 provides an overview on interestingness measures in this setting. Afterwards, Section 5 presents novel approaches for efficient subgroup discovery with numerical target concepts, that is, data structures, optimistic estimate bounds, and their integration in algorithms. The benefits of the suggested techniques are evaluated in-depth in Section 6. Finally, Section 7 concludes the paper with a summary and an outlook for future research.

2 Related Work

Mining supervised local patterns, e.g., discriminative patterns (Cheng et al. 2008), contrast sets (Bay and Pazzani 2001), emerging patterns (Dong and Li 1999) or subgroup discovery (Klösgen 1996; Wrobel 1997; Klösgen 2002; Atzmueller 2015), has been established as a versatile and effective method in data mining. While this paper focuses on subgroup discovery, recent research shows that many of these tasks differ mostly in terminology and many techniques can be transfered between tasks with little effort, cf. (Kralj Novak et al. 2009). Efficient mining algorithms can be classified in three dimensions, i.e., search strategy, data structure, and pruning mechanisms. The search strategy can directly be transfered from the binary to the numerical target case. The algorithms described here apply depth-first-search, but the proposed improvements regarding data structures and especially pruning bounds can easily be transfered to other search strategies such as Apriori (Morishita and Sese 2000; Kavšek and Lavrač 2006) or exhaustive best-first-search (Webb 1995; Zimmermann and De Raedt 2009).

Generally, numerical data can be discretized (cf. for example (Fayyad and Irani 1993; García et al. 2013)) in order to apply standard subgroup discovery for binary target concepts. One method that was specifically designed for numerical targets in subgroup discovery, is *TargetCluster* (Moreland and Truemper 2009). It uses a scoring of clustering solutions to find appropriate intervals for the target concept. Nonetheless, discretizing the target concept still leads to a loss of information. Subgroup discovery with numerical target concepts without discretization was applied in the pioneering data mining system Explora (Klösgen 1996). It applied a variety of enumeration strategies for subgroup discovery. Regarding numerical attributes, it was able to identify "mean patterns", that is, subgroups with a significantly deviating mean value in the numerical target attribute in comparison to the total population. An exemplary case study for this system using mean patterns is provided in (Klösgen 1994). That work, however, does not describe optimistic estimates or data structures that are specific for the mining of subgroups with numerical target concepts.

Measuring the interestingness of patterns is a challenging and active research topic in data mining. For most interestingness measures only binary target concepts are considered, see (Geng and Hamilton 2006) for an overview. But also for numerical target concepts a variety of different measures has been proposed. We will provide a summary on these measures in Section 3.1, also providing references to the original papers there. This includes and significantly extends the measures collected by Klösgen (2002) and Pieters et al. (2010), see also (Pieters 2010).

Optimistic estimate pruning has been recognized as a crucial method for efficient exhaustive pattern mining. This concept has originally been developed for general search algorithms (Hart et al. 1968; Webb 1995), and has later been transfered to subgroup discovery, cf. (Wrobel 1997; Grosskreutz et al. 2008). Regarding numerical target variables, Webb (2001) exploited optimistic estimates to efficiently find association rules with a numerical target variable in the rule head with a specific interestingness measure, i.e., "impact rules". By contrast, this paper presents upper bounds for a wide variety of interestingness measures. Another technique to derive optimistic estimates that is also applicable for numerical target concepts was proposed in (Morishita and Sese 2000). This approach is discussed in depth in Section 5.2.2. In order to reduce the redundancy between result patterns, generalization-aware interestingness measures have been proposed, see e.g. (Bayardo et al. 1999; Batal and Hauskrecht 2010). In previous work, we described difference-based optimistic estimates for generalization-aware measures. This novel family of optimistic estimates is also applicable to generalization-aware mean-based interestingness measures in case of numerical target concepts (Lemmerich et al. 2013). By contrast, this paper focuses on the well-established traditional type of interestingness measures that are only based on the statistics of the evaluated subgroups and assumes that redundancy reduction will be performed in a post-processing step.

The proposed two algorithms use different specialized data structures. One algorithm employs FP-Trees, cf. Han et al. (2000). These have been used before for subgroup discovery with binary target concepts, i.e., in the algorithms SD-Map (Atzmueller and Puppe 2006) and DpSubgroup (Grosskreutz 2008). In previous work, we introduced an extension of FP-trees, called generalized pattern trees (GP-trees) (Lemmerich et al. 2012), to the exceptional model mining setting, cf. (Leman et al. 2008; Atzmueller 2015), focusing on more complex target concepts. The *SD-Map** algorithm presented in this paper can be considered as a specialized version of this algorithm that additionally incorporates the computation of optimistic estimate bounds. The other algorithm, called *NumBSD*, adapts a vertical data structure based on bitsets (also called bitmaps or bitvectors), cf. also (Zaki 2000; Lemmerich et al. 2010), to numerical target concepts. A related data structure is used by the CAREN-DR algorithm (Jorge et al. 2006) to find "distribution rules". In contrast to the *NumBSD* algorithm, the bitsets are unordered and pruning is only applied with regard to the support of patterns.

Aumann and Lindell (1999, 2003) investigated a related problem setting in the context of association rules, described as *quantitative association rules*. For the discovery of desired rules, they apply a three-stage process: (1) find all frequent patterns; (2) compute an interestingness value of these patterns based on the deviation of the mean or the variance of the target; (3) filter sub-rules that are contained in more general rules. As a result, pruning is only based on the support of the patterns, in contrast to the pruning techniques proposed in this work.

Numerical data in subgroup discovery has also been investigated for the set of attributes defining the search space. In this context, the *MergeSD* algorithm, proposed in (Grosskreutz and Rüping 2009), is designed for exhaustive search. It exploits relationships between selectors of a single attribute by applying additional pruning based on a specialized data structure, the *boundTable*. Mampaey et al. (2012) analyzed the refinement step for greedy algorithms such as beam search with respect to online discretization of numerical search space attributes. They propose a method that allows for finding the best interval of a numerical attribute that is added to the current subgroup description in linear time of the number of potential cutpoints, in contrast to the quadratic time required by the trivial approach. In the field of association rule mining, related approaches have been discussed: Fukuda et al. (1996), for example, investigated numerical attributes in the rule condition of *optimized association rules*. This problem setting has been extended in (Rastogi and Shim 2002; Brin et al. 2003) to include disjunctions of intervals. In contrast to these works, this paper focuses on subgroup discovery with numerical *target* attributes.

3 Background

This section introduces the used definitions and notations. Then, the general problem setting of subgroup discovery with numerical targets is presented.

3.1 Subgroup Discovery

Subgroup discovery aims to identify patterns having the most unusual statistical characteristics with respect to the concept of interest, e. g., given by a (dependent) target variable, see e.g., (Klösgen 1996; Wrobel 1997; Klösgen 2002; Atzmueller 2015). These patterns are described by explaining (independent) variables.

A dataset $\mathcal{D} = (\mathcal{I}, \mathcal{A})$ is formally defined as an ordered pair of a set of *instances* (also called individuals, cases, or data records) $\mathcal{I} = c_1, c_2, \ldots, c_y$ and a set of attributes $\mathcal{A} = A_1, A_2, \ldots, A_z$. Each attribute $A_m : \mathcal{I} \rightarrow dom(A_m)$ is a function that indicates a characteristic of an instance by mapping it to a value in its domain. $A_m(c)$ denotes the value of the attribute A_m for the instance c. An attribute is called *nominal*, if its values are only differentiated by their name. On the other hand, an attribute A_{num} is called *numerical*, if its domain contains exclusively real valued numbers, i.e., $dom(A_{num}) \subseteq \mathbb{R}$.

A *selector* sel is a boolean function $\mathcal{I} \rightarrow \{true, false\}$ that describes a set of instances with a selection expression over one attribute. Σ denotes the set of all selectors. Typical selectors for nominal attributes are selections on single attribute values, e.g., $\text{sel}_{gender=male}$, but selectors may also contain a set of attribute values, negated values or (in case of numerical attributes) intervals. A *subgroup description* or *pattern* $P = \{\text{sel}_1, \ldots, \text{sel}_d\}, \text{sel}_i \in \Sigma, i = 1 \ldots d$, is then defined by a set of selectors which is interpreted as a conjunction, i.e., $P \mathrel{\widehat{=}} \text{sel}_1 \wedge \ldots \wedge \text{sel}_d$. We call a pattern P_{gen} a generalization of its specialization P_{spec}, iff $P_{gen} \subset P_{spec}$. For a fixed dataset, a subgroup (the extension of P) $sg(P) := \{i \in \mathcal{I} | \forall sel \in P : sel(i) = true\}$ is now given by the set of individuals that are covered by the subgroup description P. Trivially, a generalization covers all instances that are covered by its specializations: $P_{gen} \subset P_{spec} \Rightarrow sg(P_{gen}) \supseteq sg(P_{spec})$. For shorter notation, we write the number of instances covered by a pattern P as $i_P = |sg(P)|$. Consequently, $i_{\neg P}$ describes the number of instances not covered by P, and $i_\emptyset$ denotes the number of instances in the total population.

A subgroup discovery task can now be specified by a 5-tuple $(\mathcal{D}, \Sigma, T, q, k)$. $\mathcal{D}$ is the dataset. The set of all selectors Σ defines the *search space* of $2^{|\Sigma|}$ candidate subgroup descriptions in the dataset. While the construction of appropriate selectors can be a non-trivial task especially for numerical attributes, we do not focus on this problem in this paper. Instead, we consider the set of basic selectors as fixed, computed by a preprocessing step, e.g., using discretization.

The *target concept* T specifies the property of interest for the discovery task. In classical subgroup discovery, the target concept is commonly given by a certain pattern and the goal is to identify subgroups in which this target pattern occurs more/less frequently (relative to the subgroup size) than in the overall set of individuals. The value of the target concept ("target value") for an instance c is denoted by $T(c)$. For binary target concepts, we write the target share, i.e., the share of instances with a true target concept, in a subgroup (in the overall dataset) as τ_P ($\tau_\emptyset$). In many applications of subgroup discovery, the property of interest is given by a numerical attribute, see e.g., (Klösgen

and May 2002; Grosskreutz 2008; Atzmueller and Puppe 2009; Lemmerich and Atzmueller 2012; Atzmueller and Lemmerich 2013).

In general, the case of numerical target attributes can be transformed back to the binary case using discretization techniques (García et al. 2013). E.g., using the target variable *age* with $dom(age) = [0, 140]$ the group "older people" could be defined using the interval $[70, 140]$. A significant subgroup could then be formulated as *"while in the general dataset only 6% of the people are older than 70, in the subgroup described by* xy *it is 12%"*. However, such thresholds are often difficult to determine, and additional information on the distribution of the target attribute is lost. For example, a subgroup that contains many people aged between 60 and 70 will not be regarded as a subgroup containing "older people" – perhaps in contrast to some user's expectation. This discretization hides also the difference between subgroups in which the majority is around 60 years old, and those in which the majority is around 20 years old. Therefore, using the complete distribution of the numerical target attribute is potentially advantageous. The distribution of a numerical target attribute in a subgroup is more difficult to describe than the distribution of a binary target pattern: it is given by a multi-set of real values instead of just the numbers of positive and negative instances. Thus, the target distribution for a subgroup description P is often compared in terms of one or more distributional properties, e.g., the mean value μ_P, the median med_P or the variance σ_P^2 of the numerical target attribute. For example, an interesting subgroup based on the mean values can be formulated as: *"While in the general dataset the mean age is* 56 *years, in the subgroup described by* xy *it is* 62 *years"*.

Given a database $\mathcal{D}$ and a target concept T, the interestingness measure $q : 2^{\Sigma} \rightarrow \mathbb{R}$ maps every pattern in the search space to a real number that reflects the *interestingness of a pattern*. To keep this paper concise, we focus on traditional interestingness measures which are purely dependent on the coverage of the subgroup, so $q : 2^{\mathcal{I}} \rightarrow \mathbb{R}$. The authors are aware that this significantly restricts the scope of this paper, as more recent variations that also include the description of the subgroup in the selection process, see (Bayardo 1998; Atzmueller et al. 2009; Batal and Hauskrecht 2010; Lemmerich and Puppe 2011), are left out. For simpler notations, $q(sg(P))$ and $q(P)$ are used equivalently. A popular family of interestingness measures for binary targets are the Klösgen measures, which trade off the coverage of a subgroup with the deviation of its target share: $q^a_{Kl}(P) = {i_P}^a \cdot (\tau_P - \tau_\emptyset), a \in [0, 1]$ (Klösgen 1996), see also for example (Wrobel 1997; Lavrač et al. 2004; Grosskreutz and Rüping 2009; Atzmueller 2015). Interestingness measures for numerical target concepts will be discussed in detail in Section 4. Interestingness measures imply an ordering of the subgroups in the search space. Two interestingness measures $q_1(P)$ and $q_2(P)$ that imply the identical order for any pair of subgroups in a dataset are called *order equivalent*, denoted as $q_1(P) \sim q_2(P)$. Obviously, order equivalent interestingness measures lead to identical results in an exhaustive top-k search.

Finally, the integer k gives the number of returned patterns of this task. The result of a subgroup discovery task is the set of k subgroup descriptions

with the highest interestingness values according to the chosen interestingness measure. Note that even if the score of a subgroup is only determined by the subset that is covered by a subgroup description the result of the discovery algorithm is still a pattern with a description that is interpretable by humans.

4 Interestingness Measures for Numerical Target Concepts

This section presents a concise and comprehensive survey on interestingness measures for numerical target concepts, substantially extending previous discussions, see for example (Klösgen 1996, 2002; Pieters et al. 2010).

For numerical target concepts, many interestingness measures extract certain data characteristics, e.g., the mean or the median value, from the respective dataset and compare those values obtained in the subgroup and in the overall dataset. We categorize the interestingness measures for numerical target concepts with respect to the used data characteristics:

1. *Mean-based interestingness measures:* A simple approach to score subgroups is to compare the mean value in the subgroup μ_P with the mean value in the overall dataset $\mu_\emptyset$. A pattern is then considered as interesting, if the mean of the target values is (significantly) higher within the subgroup. In that direction, several interestingness measures have been proposed:
   (a) *Generic mean-based functions:* A generic formalization for a variety of such measures can be constructed by adapting the Klösgen interestingness measures for binary targets: the target shares $\tau_P, \tau_\emptyset$ of subgroups and the general dataset are replaced by the respective mean values of the target variable in the subgroup μ_P and in the overall dataset $\mu_\emptyset$. This results in: $q^a_{mean}(P) = {i_P}^a \cdot (\mu_P - \mu_\emptyset), a \in [0, 1]$. Higher values of a favor larger subgroups, lower values favor larger deviations in the target share. These measures include the Klösgen measures for binary targets as a special case, if the binary target concept is interpreted as an indicator function ($T(c) = 1$ for true target concepts, $T(c) = 0$, otherwise), since the mean values in the formula are then equal to the respective target shares.
   This generic family of functions is either equal or order equivalent to several other interestingness measures proposed in literature, such as the *average function* for $a = 0$, *mean test* (Grosskreutz 2008) and *z-score* (Pieters et al. 2010) for $a = 0.5$, or *impact* (Webb 2001) for $a = 1$.
   (b) *Generic symmetric mean-based functions:* To discover subgroups with decreased as well as increased target values in a single run of the discovery algorithm, the difference of the target shares can be replaced by the respective absolute value: $q^a_{sym}(P) = {i_P}^a \cdot |\mu_P - \mu_\emptyset|$. As an alternative, we can also use the squared difference instead: $q^a_{sq}(P) = {i_P}^a \cdot (\mu_P - \mu_\emptyset)^2$. This results in measures that are order equivalent to $q^{\frac{a}{2}}_{sym}$.
   (c) *Variance reduction:* Another symmetric measure, which has been introduced in the context of regression tree learning, is the variance reduction (Breiman et al. 1984; Klösgen 1996): $q_{vr}(P) = \frac{i_P}{i_\emptyset - i_P} \cdot (\mu_P - \mu_\emptyset)^2$

(d) *Interclass variance:* The interclass variance was proposed to measure the correlation between a pattern and a numerical target attribute: $q_{iv}(P) = i_P \cdot (\mu_P - \mu_\emptyset)^2 + i_{\neg P} \cdot (\mu_{\neg P} - \mu_\emptyset)^2$, cf. (Morishita 1998; Morishita and Sese 2000)

2. *Variance-based measures:* Aumann and Lindell (2003) proposed to identify patterns with an unusual variance.
   (a) *Generic variance-based functions:* This can be accomplished by replacing the target shares $\tau_P, \tau_\emptyset$ with standard deviations $\sigma_P, \sigma_\emptyset$ in the Klösgen measures, resulting in: $q^a_{sd}(P) = {i_P}^a \cdot (\sigma_P - \sigma_\emptyset), a \in [0, 1]$. As before, this allows – but also requires – controlling the coverage of the results using the size parameter a. These measures do not directly correspond to a statistical significance test. Aumann and Lindell propose to use an F-Test (Aumann and Lindell 2003) for testing statistical significance, but this test should be applied carefully due to its strong sensitivity to the non-normality of the distribution (Box 1953).
   (b) *t-score:* The t-score $q_t(P) = \frac{\sqrt{i_P} \cdot (\mu_P - \mu_\emptyset)}{\sigma_P}$ (Pieters et al. 2010; Klösgen 2002) incorporates the mean μ_P and the standard deviation σ_P of the target values in a subgroup P. It reflects the significance of the deviation of target values in a subgroup using a Student's t-test. However, a direct statistical interpretation of the t-score should be avoided if the target concept is not normally distributed and the subgroup size is small, e.g., $i_P < 30$.
3. *Median-based measures:* Statistics based on the mean target value of subgroups are known to be sensitive to outliers. Therefore, it can be favorable to use the more stable median instead of the mean value.
   (a) *Generic median-based measure:* A generic family of median-based interestingness measures can again be derived by a small adaptation of q^a_{mean}: $q^a_{med}(P) = {i_P}^a \cdot (med_P - med_\emptyset)$, where med_P is the median of target values in the subgroup and $med_\emptyset$ the median in the total population. In general, there is no direct interpretation of these measures with respect to a statistical significance test.
   (b) *Median χ^2 Test:* As proposed in (Pieters et al. 2010), the significance of a χ^2-test that uses the median of the target attribute in the total population as a discretization cut-point can be applied as an interestingness measure. From a computational point of view, this is accomplished by performing discretization as a pre-processing step and running a subgroup discovery algorithm for binary targets. Therefore, this measure will not be discussed with respect to efficient mining in this work.
4. *Rank-based measures:* A variety of statistical tests for the deviation of numerical variables use the ranks of the target attribute instead of the target values themselves. That is, the instance with the highest target value is mapped to rank one, the instance with the second highest target value is mapped to rank two, and so on. This reduces the sensitivity to outliers compared to mean-based tests. Additionally, rank-based methods can also be applied to ordinal attributes.

(a) *Mann-Whitney measure:* Klösgen (1996) and later Pieters et al. (2010) proposed an interestingness measure based on the Mann-Whitney (also Wilcoxon-Mann-Whitney) rank sum test on statistical significance. The measure compares the difference of the mean of ranks in the subgroup with the overall mean of ranks and computes its significance using a z-statistic. It is defined as:

$$q_{mw}(P) = i_P \cdot \frac{\frac{\mathcal{R}_P}{i_P} - \frac{i_\emptyset + 1}{2}}{\sqrt{\frac{i_P i_{\neg P}(i_\emptyset + 1)}{12}}} \sim \sqrt{\frac{i_P}{i_{\neg P}}} \cdot \left(\frac{\mathcal{R}_P}{i_P} - \frac{i_\emptyset + 1}{2} \right) := q_{mw'}(P),$$

where $\mathcal{R}_P$ is the sum of ranks within the subgroup P.

(b) *AUC measure:* This interestingness measure proposed in (Pieters et al. 2010) determines the area under the ROC curve (Fawcett 2006). It can be computed as:

$$q_{auc}(P) = \frac{\mathcal{R}_{\neg P} - \frac{i_{\neg P} \cdot (i_{\neg P} + 1)}{2}}{i_P \cdot i_{\neg P}},$$

with $\mathcal{R}_{\neg P}$ being the sum of ranks in the complement of the subgroup P. This measure is independent of the subgroup's coverage.

5. *(Full) Distribution-based measure (Kolmogorov-Smirnov measure):* The significance according to a Kolmogorov-Smirnov statistical test has been proposed to discover so called *distribution rules* (Lucas et al. 2007; Jorge et al. 2006). The measure is order equivalent to the test statistic of this test: $q_{ks}(P) = \sqrt{\frac{i_P \cdot i_{\neg P}}{i_\emptyset}} \Delta_{(P,\neg P)}$, where $\Delta_{(P,\neg P)}$ is the supremum of differences in the empirical distribution function induced by the subgroup P and its complement $\neg P$. The empirical distribution function is a function that computes for each value v in the target attribute's domain the fraction of pattern instances with a target value smaller or equal to v. This measure can capture increases as well as decreases of the target values.

Answering the question, when to use which measure, is left to future work. Instead, we concentrate in this paper on the efficient mining with the presented measures.

5 Efficient Exhaustive Approaches for Subgroup Discovery with Numerical Properties of Interest

In this section, we investigate efficient subgroup discovery with numerical target concepts. We first discuss the adaptation of data structures as well as options for optimistic estimate pruning for the presented interestingness measures. After that, we present two algorithms that incorporate these approaches.

5.1 Data Representations

Specialized data structures allow for the efficient computation of subgroup statistics required by interestingness measures and optimistic estimate bounds. Below, we introduce adaptations of two data structures to the setting of subgroup discovery with numerical target concepts, that is, FP-trees and bitset-based data structures. We primarily focus on generic mean-based interestingness measures and outline adaptations for other measures only briefly.

5.1.1 Adaptations of FP-trees

FP-trees (Han et al. 2000) have been proposed as efficient data structures for the mining of frequent itemsets. These extended prefix tree structures store the relevant information in a compressed way. Each tree node contains a reference to a selector and a frequency count. Additionally, links between nodes referring to the same selector are maintained. An FP-tree is built in two passes over the dataset instances: the initial pass sorts the selectors according to their frequency in the dataset. In the second pass, data instances are inserted one-by-one into the FP-tree. The order of the selectors increases the chance of shared prefixes between data instances, thus decreasing the overall size of the FP-tree. The resulting FP-tree contains the complete condensed frequency information for each pattern. A mining algorithm starts with creating an FP-tree for the initial dataset. Patterns containing exactly one selector are evaluated by the frequencies collected during the first pass over the dataset. Then, the algorithm recursively extends those patterns by adding further selectors in a depth-first manner, building conditional trees conditioned on the current pattern prefix. In this way, compact and efficient mining of the condensed tree structure is enabled. For more detailed information on FP-trees in general we refer to (Han et al. 2000, 2004). In previous work, we have shown how FP-trees can be transfered to subgroup discovery with binary targets (Atzmueller and Puppe 2006) and to the exceptional model mining setting (Lemmerich et al. 2012). These approaches, however, did not incorporate optimistic estimate bounds.

FP-trees consist of nodes, which are connected by two link structures, tree links and auxiliary links. Modifications for subgroup discovery do not affect the link structures, but extend the information that is stored in each node. In the case of a binary target, an FP-tree node contains information on the number of positive and negative instances for the respective instance set. In the case of a numerical target concept and a mean-based interestingness measure, the sum of the target values and the instance count is stored instead. This enables the computation of the mean target value of an instance set and thus allows for determining the interestingness value. Using these adaptations for numerical target variables, the case of a binary target is included as a special case, if the value of the target is set to 1 for true target concepts, or set to 0 for false target concepts, respectively.

In contrast to the binary case, additional information is required to efficiently compute the optimistic estimates for numerical target concepts that

Table 1 A toy dataset used in the illustrating examples below.

Instance	Target value	sel_A	sel_B	sel_C
c_1	100	$true$	$false$	$false$
c_2	75	$true$	$false$	$false$
c_3	60	$false$	$true$	$false$
c_4	53	$true$	$true$	$true$
c_5	40	$false$	$false$	$false$
c_6	35	$false$	$false$	$false$
c_7	25	$false$	$true$	$false$
c_8	12	$true$	$false$	$false$

will be introduced in Section 5.2.4. This information can also be effectively stored in the tree nodes: to compute the optimistic estimate oe^1_{mean}, that will be presented in Theorem 10, one additional field is used, which is initialized with 0. For each instance c corresponding to the node, the value $\max(0, T(c) - \mu_\emptyset)$ is added to this field. Like the other stored values, this field is then propagated recursively, when conditional trees are built. In doing so, the value stored in this field of each node reflects the sum $\sum_{c \in P: T(c) > \mu_\emptyset} (T(c) - \mu_\emptyset)$, that is, the exact value of the optimistic estimate. Thus, the optimistic estimate is directly available if pruning options are checked. Analogously, for the optimistic estimate $\overline{oe}^a_{mean}(P) = \widetilde{p}_P{}^a \cdot (T_P^{max} - \mu_\emptyset)$, see Theorem 12, each node must keep track of the number of instances $\widetilde{p}$ that have a target value greater than the population mean target value, and the maximum target value corresponding to this node T^{max}. These are propagated accordingly and allow for the efficient computation of these bounds.

Adaptations for other interestingness measures For other interestingness measures, different kinds of information need to be captured in the tree nodes. To apply a variance-based interestingness measure such as the t-score measure $q_t(P)$, in addition to the sum of values the sum of squared values needs to be stored to determine the variance within the subgroup, cf. (Lemmerich et al. 2012). Unfortunately, it is difficult to determine optimistic estimates for this function in general, see Section 5.2.

To compute the symmetric generic mean-based measures, no additional information is required other than the sum of values and the frequency count of instances. In order to determine optimistic estimates, it is required to additionally store the sum of target values that are below the population target mean, and the minimum target value. Similarly, for the variance reduction $q_{vr}(P)$ the instance count and the overall sum of target values are required for computing the interestingness itself. For the computation of optimistic estimate bounds, the sum of target values higher, resp. lower than the population mean as well as the minimum and maximum target value are then also required.

The generic median-based measures cannot be computed by applying an FP-tree-based data structure since more than one pass over the subgroup is required to compute the median, cf. (Lemmerich et al. 2012).

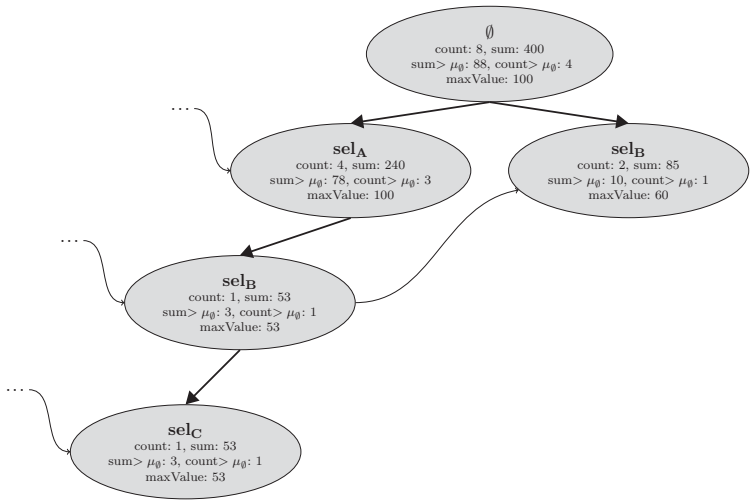

Fig. 1 The initial FP-tree (without header nodes) for the example toy dataset. The information stored in the nodes depends on the used interestingness measure. For generic mean-based measures, each node stores the node count, the sum of target values, the sum of target values above the population's target mean, the number of instances with a target value above the target mean and the maximum target value.

For computing rank-based interestingness measures using a variation of FP-trees, the ranks of instances must be determined in a preprocessing step. Afterwards, ranks replace the original target values in the algorithm itself. In the FP-tree nodes, only the instance count and the sum of ranks is stored and aggregated. For computing optimistic estimates for the Mann-Whitney measure, see Theorem 17, additionally the sum of ranks above the population's mean rank, the maximum rank, and the number of instances with a rank higher than the population's mean rank need to be stored.

Example 1 Consider the toy dataset with 8 instances and 3 selectors in the search space shown in Table 1. The initial FP-tree for this dataset (without header nodes) is depicted in Figure 1. The information stored in the nodes depends on the used interestingness measure. For generic mean-based measures, each node stores the node count, the sum of target values, the sum of target values above the population's target mean, the number of instances with a target value above the population's target mean and the maximum target value.

5.1.2 Adaptation of Bitset-based Data Structures

Vertical data representations such as bitsets, cf. (Klösgen 1995; Lemmerich et al. 2010), are an alternative to the FP-tree data structures discussed above. Here, the instances that correspond to a subgroup pattern are stored in words of single bits. Each word contains as many bits as the dataset contains cases.

The *i-th* bit in each word belongs to the *i-th* instance of the dataset. The bit is set to 1 if this instance is covered by the respective subgroup description, and is set to 0 otherwise.

For each selector in the search space one such bitset is generated. To create bitsets that correspond to conjunctive patterns a logical AND operation is performed on the bitsets of the involved selectors. The size of a subgroup can then be derived by determining the cardinality of the bitset, that is, the number of bits set to 1. For the efficient counting of bits in a bitset which are set to *true*, specialized algorithms and even supporting hardware implementations have been developed, see for example (El-Qawasmeh 2003).

For subgroup discovery with binary targets, one additional bitset reflects the occurrence of the target concept. To adapt bitset-based, vertical data structures to numerical target settings and to the introduced ordering-based bounds, see Section 5.2.2, two adaptations to the data structure in the binary setting are necessary. First, the instances of the total population are initially sorted in descending order with respect to the target variable. The ordering allows for an easy computation of ordering-based optimistic estimate bounds. Second, the numerical target values are stored in an additional array in descending order. This replaces the additional bitset used for the target concept in the binary case. The array of target values and the bitsets for the selectors correspond to each other via the position of the instances, i.e., the target value of an instance, which is represented by the n-th bit of a bitset, is given at position n of the array of target values. The computation of (for example) the mean value of a subgroup $sel_A \wedge sel_B$, requires one iteration over all bits that are set to true in the bitset that corresponds to this subgroup. For each bit that is set to true the respective target value of the array is added to a total sum and the count of instances is incremented. These statistics are then used to compute the mean value. Internally, each bitset is divided into words (e.g., of 32 or 64 bits), on which logical boolean operations (such as *OR* and *AND*) can be applied very efficiently. The fast bounds presented in Section 5.2.3 can be checked after each word. The construction of the bitsets and the target value array is accomplished in one single pass through the database. The rest of the algorithm can then operate exclusively on the generated data structure representation.

Target:	100	75	60	53	40	35	25	12
sel_A:	1	1	0	1	0	0	0	1
sel_B:	0	0	1	1	0	0	1	0
sel_C:	0	0	0	1	0	0	0	0

Fig. 2 The adapted bitset-based representation for the example dataset from Table 1.

Example 2 As a simple example, the adapted bitset-based data structure for the example dataset of Table 1 is shown in Figure 2.

5.2 Optimistic Estimates

Given an interestingness measure and a subgroup, an optimistic estimate is an upper bound for the interestingness score of all specializations of this subgroup. It is used to speed up subgroup discovery algorithms: if the optimistic estimate of a subgroup is lower than the interestingness value required in the result set, then the specializations of this subgroup can be excluded from the search. As shown for binary target concepts, e. g., (Wrobel 1997; Grosskreutz et al. 2008) this can reduce the number of candidate subgroups in discovery algorithms by orders of magnitude. Nonetheless, for subgroup discovery with numerical target concepts optimistic estimate bounds have only received limited attention in the literature so far.

This section thoroughly analyzes optimistic estimate pruning for subgroup discovery with numerical target concepts for a large variety of interestingness measures. After a short formal definition of optimistic estimates, we discuss an approach for deriving optimistic estimates for convex interestingness measures that has been proposed by Morishita and Sese (2000). Extending this direction of research, we introduce the formalism of *ordering-based optimistic estimate bounds* as a tool that allows for deriving optimistic estimates. With this formalism, optimistic estimates can be determined easily for a wider range of interestingness measures, including previously unbounded measures. This is demonstrated for several types of interestingness measures. Afterwards, we present a novel technique that provides a series of increasingly tighter upper bounds, which is computed by using only a part of the instances covered by a subgroup. Additionally, optimistic estimates in closed-form expressions are derived for the discussed interestingness measures that can be used also in combination with FP-tree-based data representations.

5.2.1 Formal Definition

Given an interestingness measure $q(P)$, an optimistic estimate oe_q is a function $2^{\Sigma} \rightarrow \mathbb{R}$, such that the interestingness score of each refinement, i.e., a subset of the current subgroup, is lower or equal to the function value for this subgroup:

$$\forall S \supset P : q(S) \leq oe_q(P)$$

For efficient mining, more precise bounds are beneficial, since they allow for pruning larger parts of the search space. *Tight optimistic estimates* are "estimates that are as conservative as possible with respect to the information at hand" (Grosskreutz 2008). In our scenario, we call an optimistic estimate tight if a subgroup contains a subset of instances that attains the interestingness score given by the optimistic estimate.

$$\exists r \subseteq sg(P) : q(r) = oe(P)$$

Tight optimistic estimates are the most precise optimistic estimates that can be derived from the distribution of the target concept in the pattern alone.

5.2.2 Ordering-based Bounds

Morishita and Sese (2000) proposed a general scheme for determining optimistic estimates for a specific class of interestingness measures, that is, measures that are *convex* in a certain space of dimensions defined by subgroup statistics. In that approach, each subgroup is mapped to a so-called *stamp point* in the respective space according to its respective statistics. Then, the interestingness measure maps each of these points to a score. If this function is convex, it can be shown that its maximum values within a convex polygon are attained at the vertices of the polygon.

Regarding numerical target concepts, the authors investigate the interclass variance which is a convex function in the two-dimensional parameter space $(\Sigma T(c), i_p)$ constructed using the sum of the target values $\Sigma T(c)$ and the number of subgroup instances i_p as its dimensions. A bounding convex polygon for all points in the space that correspond to the specializations of a subgroup P can be constructed as follows: the instances of the dataset are sorted according to their target values. Then, each split point in the target values is considered. That is, for each different target value, the refinement that contains all subgroup instances with target values higher than the split point and the refinement of all instances lower than the split point is evaluated. Since the stamp points of these refinements form a convex polygon that contains all specializations of P, the maximum interestingness score of the evaluated refinements is an optimistic estimate for P.

In this section, we show that a similar approach is more generally applicable and can be used for many (but not all) interestingness measures including measures that are not convex or are not even functions in the $(\Sigma T(c), i_p)$ space. We start with a formal definition of the desired property.

Definition 1 Let s_j^{desc} (s_j^{asc}) be the set of instances that consists of the j instances of $sg(P)$ with the highest (lowest) target values. Then, an interestingness measure q is *one-pass estimable by ordering*, if it holds for any subgroup P and any refinement $r \subseteq sg(P)$, that

$$q(r) \leq \max\left(q(s_1^{desc}), \ldots, q(s_{i_P}^{desc})\right).$$

An interestingness measure q is *two-pass estimable by ordering*, if it holds for any subgroup P and any refinement $r \subseteq sg(P)$ that

$$q(r) \leq \max\left(q(s_1^{desc}), \ldots, q(s_{i_P}^{desc}),\ q(s_1^{asc}), \ldots, q(s_{i_P}^{asc})\right).$$

In other words, for measures that are one-pass estimable by ordering, the interestingness score never decreases if one of the instances is exchanged with another instance that has a greater target value. For such measures, only i_P candidates must be considered to find the best refinement of a subgroup P.

This motivates the following approach for subgroup discovery with numerical target concepts using such interestingness measures: in a preprocessing step, the instances in the database are sorted in descending order with respect to their target values. Whenever a subgroup is evaluated, instances are added

one by one to the subgroup, starting with the one with the highest target value. After each addition, the interestingness of the instance set is evaluated. The maximum of these interestingness values is used as a tight optimistic estimate. In doing so, only a single pass over each subgroup is required. For measures that are two-pass estimable by ordering, the best subset of instances is found by traversing the current subgroup twice — once in descending and once in ascending order of the target values. In both passes, instances are added one by one to the current set of instances. The overall optimistic estimate is then given by the maximum of all those scores. In doing so, $2 \cdot i_P$ subsets of the instances of the current subgroup are considered as candidates to find its best refinement. As shown in (Morishita and Sese 2000), measures that are convex in the $(\Sigma T(c), i_p)$ space are two-pass estimable by ordering. The more general property defined here is not only more generally applicable, but in the authors' opinion also more convenient to prove.

First, we investigate the generic mean-based interestingness measure q^a_{mean}. While these measures are convex in the $(\Sigma T(c), i_p)$ space for the special case of $a = 1$, they are not convex for the general case with arbitrary a, see the appendix for a proof. Nonetheless, the generic mean-based interestingness measures are one-pass estimable by ordering.

Theorem 1 *The interestingness measures $q^a_{mean}(P) = {i_P}^a \cdot (\mu_P - \mu_\emptyset)$ are one-pass estimable by ordering.*

Proof We consider any refinement $r \subseteq sg(P)$ and compare it to the instance set $r^* = s^{desc}_{|r|}$. This is the subset with the same number of covered instances as r, but the highest target values contained in $sg(P)$. Then, $|r^*| = |r|$ and $\mu_{r^*} \geq \mu_r$. It follows that according to a mean-based interestingness measure the refinement r^* is at least as interesting as r: $q^a_{mean}(r) = |r|^a \cdot (\mu_r - \mu_\emptyset) \leq |r^*|^a \cdot (\mu_{r^*} - \mu_\emptyset) = q^a_{mean}(r^*)$ □

Example 3 Consider the pattern P_A for the selector sel_A in the toy dataset of Table 1. Here, the population mean is $\mu_\emptyset = 50$ and the target values for the subgroup instances c_1, c_2, c_4, and c_8 are $100, 75, 53$, and 12. These instances are sorted in descending order according to their target values (as already done in this case) to generate the subsets s^{desc}_j. For example, the subset s^{desc}_3 contains the 3 instances c_1, c_2, and c_4. For each of the subsets $s^{desc}_i, i = 1 \ldots 4$, the score of the interestingness measure is computed. In this example, the mean test measure $q^{0.5}_{mean}(P) = \sqrt{i_P} \cdot (\mu_P - \mu_\emptyset)$ is used. The resulting scores are: $q^{0.5}_{mean}(s^{desc}_1) = \sqrt{1}(100 - 50) = 50$; $q^{0.5}_{mean}(s^{desc}_2) = \sqrt{2}(87.5 - 50) \approx 53$; $q^{0.5}_{mean}(s^{desc}_3) = \sqrt{3}(76 - 50) \approx 45$; $q^{0.5}_{mean}(s^{desc}_3) = \sqrt{4}(60 - 50) = 20$. The maximum of these interestingness scores, in this case $q^{0.5}_{mean}(s^{desc}_2) \approx 53$, then defines an optimistic estimate bound for the subgroup P_A.

Theorem 2 *The median-based interestingness measures $q^a_{med}(P) = {i_P}^a \cdot (med_P - med_\emptyset)$ are one-pass estimable by ordering.*

Proof Analogously to Theorem 1, replacing the mean with the median. □

Note, that median-based measures cannot be considered as functions in the $(\Sigma T(c), i_p)$ space at all.

Theorem 3 *The rank-based interestingness measures $q_{mw}(P)$ and $-q_{auc}(P)$ are one-pass estimable by ordering.*

Proof For any refinement r with $|r| = j$, the sum of the ranks gets maximized (or minimized, depending on the ordering of the ranking) for s_j^{desc}, therefore $q(s_j^{desc}) \geq q(r)$ for these interestingness measures. □

The symmetric mean-based measures are two-pass estimable by ordering. This could also be shown by the convexity of these functions. However, proving the property directly is much simpler and more concise.

Theorem 4 *The symmetric mean-based measures $q_{sym}^a(P) = {i_p}^a \cdot |\mu_P - \mu_\emptyset|$ are two-pass estimable by ordering.*

Proof Consider any refinement $r \subseteq sg(P)$. Without loss of generality, let $j = |r|$ be the number of instances covered by r. If $\mu_r \geq \mu_\emptyset$ then $q_a(s_j^{desc}) \geq q_a(r)$ in analogy to the proof of Theorem 1. Otherwise, $\mu_r < \mu_\emptyset$ and we can conclude that $q_a(s_j^{asc}) \geq q_a(r)$ since $|s_j^{asc}| = |r|$ and $\mu_{s_i^{asc}} \leq \mu_r$ and therefore $|\mu_{s_j^{asc}} - \mu_\emptyset| \geq |\mu_r - \mu_\emptyset|$. □

Theorem 5 *The interestingness measure variance reduction $q_{vr}(P) = \frac{i_P}{i_\emptyset - i_P} \cdot (\mu_P - \mu_\emptyset)^2$ is two-pass estimable by ordering.*

Proof For any refinement $r \subseteq sg(P)$ that covers $j = |r|$ instances, it holds that $(\mu_P - \mu_\emptyset)^2 \leq \max\left((\mu_{s_j^{desc}} - \mu_\emptyset)^2, (\mu_{s_j^{asc}} - \mu_\emptyset)^2\right)$. Thus, we can conclude that $q_{vr}(r) \leq \max\left(q_{vr}(s_j^{desc}), q_{vr}(s_j^{asc})\right)$. □

Theorem 6 *The interestingness measure interclass variance $q_{iv}(P) = i_P \cdot (\mu_P - \mu_\emptyset)^2 + i_{\neg P} \cdot (\mu_{\neg P} - \mu_\emptyset)^2$ is two-pass estimable by ordering.*

Proof Shown in Morishita and Sese (2000) by the convexity of the measure. □

Note that due to the symmetry of this measure between the subgroup and its complement, further optimizations in the implementation are possible.

Although the estimability by ordering seems like an intuitive property for interestingness measures, it does not hold for all measures:

Theorem 7 *The generic variance interestingness measure $q_{sd}^a(P) = {i_P}^a \cdot (\sigma_P - \sigma_\emptyset)$ is* not *one-pass or two-pass estimable by ordering.*

Proof Assume that the standard deviation in a dataset is overall $\sigma_\emptyset = 1$ and the subgroup P_4 covers 3 instances with target values $T(c_1) = 10$, $T(c_2) = 0$, and $T(c_3) = -10$. The subset r^* with highest standard deviation and therefore the highest score according to $q_{sd}^0(P)$ then consists of the two instances c_1 and

c_3. The mean value for this subset is $\mu_{r^*} = 0$ and its variance is $sd^2_{r^*} = \frac{(10-0)^2+(0-(-10))^2}{2-1} = 200$. The interestingness score for this subset is then $q^0_{sd}(r^*) = 2^0 \cdot 200 - 1 = 199$, which is greater than the interestingness score of all subsets s^{desc}_j or s^{asc}_j. □

Theorem 8 *The t-score interestingness measure* $q_t(P) = \frac{\sqrt{i_P}\cdot(\mu_P - \mu_\emptyset)}{\sigma_P}$ *is* not *one-pass or two-pass estimable by ordering.*

Proof Consider a dataset with $\mu_\emptyset = 0$ and a subgroup P within the dataset that contains four instances $i_1, \ldots, i_4$ with the target values $T(i_1) = 20$, $T(i_2) = 10$, $T(i_3) = 10 + \epsilon$ with $0 < \epsilon \ll 0.1$, and $T(i_4) = 0$. Then the best refinement r^* contains the instances i_2 and i_3. The interestingness of r^* approaches infinity if ϵ approaches 0. Thus, $q_t(r^*) > max(q_t(s^{desc}_i), q_t(s^{asc}_i))$ for any i given a small ϵ. This contradicts the definition of one-pass and two-pass estimability by ordering. □

The novel algorithm *NumBSD*, which is introduced in Section 5.3, incorporates the presented ordering-based estimates in an efficient algorithm.

5.2.3 Fast Bounds using Limited Information

This section presents a novel method to speed up the computation process by applying a sequence of less tight upper bounds, which are computed during the evaluation of a single subgroup. The bounds are determined only using the refinements s^{desc}_j of P, that is, the j instances in P with the highest target values. For that, we focus exclusively on the generic mean-based interestingness measures q^a_{mean}.

The main idea is as follows: as in the previous section, instances are considered for subgroup evaluation one-by-one in descending order of the target values. After each instance, the interestingness score for the set s^{desc}_j of the already incorporated j instances is computed. By definition, the maximum of all interestingness values $q(s^{desc}_j)$ is an optimistic estimate for interestingness measures that are one-pass estimable by ordering. We now consider a certain point in time during this pass over the dataset, at which the first n instances have already been processed. At this point, it is guaranteed that the target values for all subsequent instances in the subgroup are not greater than the target value of the current case. This fact is used to determine an upper bound for all the interestingness values $q(s^{desc}_j)$ that have not yet been computed: for all instance that have not been visited yet it is assumed that the target value of these instances is equal to the target value of the instance, which was added last. By assuming larger target values, the computed interestingness value according to any generic mean-based interestingness measure q^a_{mean} always increases, since the interestingness measure is one-pass estimable. Thus, we compute the maximum value that is obtained by adding any number of instances with the current target value. This forms an upper bound for the remaining interestingness values $q(s^{desc}_j)$, $j > n$. If this less tight upper bound

already indicates that none of the refinements of the current subgroup (nor the current subgroup itself) will be added to the result set, then we can skip the rest of the evaluation of this subgroup. More formally, we capture this approach using the following theorem:

Theorem 9 *For a subgroup P, let $s_n^{desc} \subseteq sg(P)$ be the n instances with the highest target values. Furthermore, let $\sigma = \sum_{c \in s_n^{desc}} T(c)$ be the sum of target values for s_n^{desc} and θ the lowest target value for the instances in s_n^{desc}, that is, the n^{th}-highest target value in $sg(P)$. Then, an optimistic estimate for generic mean-based interestingness measures $q_{mean}^a(P) = {i_P}^a \cdot (\mu_P - \mu_\emptyset)$ is given by:*

$$oe_{fast}^a(P) = \max\left(q_{mean}^a(s_1^{desc}), \ldots, q_{mean}^a(s_n^{desc}), oe_{remaining}^a(P)\right),$$
$$oe_{remaining}^a(P,n) = {i_P}^a \cdot \left(\frac{\sigma + (i_P - n)\cdot\theta}{i_P} - \mu_\emptyset\right)$$

Proof The theorem is proven by showing that for each subset $r \subseteq sg(P)$ the interestingness score is not higher than the provided optimistic estimate: $q_{mean}^a(P) \leq oe_{fast}^a(P)$. Since $q_{mean}^a(P)$ is one-pass estimable, for each r the interestingness is lower than the interestingness of the subset that has the same number of instances but covers the instances with the highest target values: $q_{mean}^a(r) \leq q_{mean}^a(s_{|r|}^{desc})$. Thus, the theorem holds for all r with $|r| \leq n$.

For all refinements r with $|r| > n$, it remains to show that $q_{mean}^a(s_{|r|}^{desc}) \leq oe_{fast}^a(P)$. To do so, the interestingness of s_j^{desc} for any $j = n + x$ with $x > 0, x \leq x_{max}, x_{max} = i_P - n$ is estimated. Let c_k be the instance with the k$^{\text{th}}$-highest target value. Then, it holds that:

$$\begin{aligned} q_{mean}^a(s_j^{desc})(P) &= j^a \cdot \left(\frac{\sum_{k=1}^{j} T(c_k)}{j} - \mu_\emptyset\right) \\ &= (n+x)^a \cdot \left(\frac{\sum_{k=1}^{n} T(c_k) + \sum_{k=n+1}^{x} T(c_k)}{n+x} - \mu_\emptyset\right) \\ &\leq (n+x)^a \cdot \left(\frac{\sigma + x\cdot\theta}{n+x} - \mu_\emptyset\right) := f^a(x) \end{aligned}$$

This inequality is based on the fact that the target values are in descending order and it therefore holds for $k > n$ that $T(c_k) \leq T(c_n) = \theta$. The function $f^a(x)$ describes an upper bound for the interestingness of the instance set s_j^{desc} that consists of x more instances than the last evaluated instance set s_n^{desc}. Unfortunately, the size j of the instance set with the maximum interestingness and the respective x-value are not known. However, an upper bound for the interestingness of $q(s_j^{desc})(P)$ is given by the maximum value of $f^a(x)$. For this family of functions, it can be shown by computing the first and second derivative that the maximum value is reached either at $x = 0$ or at $x = x_{max} = i_P - n$. The formal proof is provided in the appendix of this paper. This means that the maximum upper bound is reached, if either none or all

remaining instances are added to the last evaluated instance with an assumed target value of θ. For $x = 0$, the value of the function f^a is equal to the interestingness score of the instance set s_n^{desc}: $f^a(0) = (n+0)^a \cdot \left(\frac{\sigma+0\cdot\theta}{n+0} - \mu_\emptyset\right) = n^a \cdot (\mu_{s_n^{desc}} - \mu_P) = q^a_{mean}(s_n^{desc})$.

As a consequence it holds for all $r \subseteq sg(P)$ with $|r| = j > n$ that

$$\begin{aligned} q^a_{mean}(r) &\leq q(s_j^{desc}) \\ &\leq f^a(x) \\ &\leq \max\left(f^a(0), f^a(i_P - n)\right) \\ &= \max\left(q^a_{mean}(s_n^{desc}), {i_P}^a \cdot \left(\frac{\sigma + (i_P - n)\cdot\theta}{i_P} - \mu_\emptyset\right)\right) \\ &\leq oe^a_{fast}(P) \end{aligned}$$

This proves the theorem. □

This theorem provides an upper bound for the interestingness score of all specializations of a subgroup and also for the interestingness of the subgroup itself. It is based only on a subset of the instances of a subgroup, that is, the ones with the highest target values. Thus, the above estimates can be checked during an iteration over the subgroup instances even before this iteration has finished. If after only a few instances the computed upper bound indicates that the subgroup itself and all its specializations do not have a sufficient interestingness for the result set, then the evaluation of the subgroup can be stopped. In doing so the majority of the subgroup instances does not need to be considered, thus speeding up the subgroup discovery process. To the authors' knowledge, this is the first approach that uses optimistic estimates that are based only on a part of a subgroup's instances.

Example 4 Again, consider the subgroup P_A for the selector sel_A in the toy dataset of Table 1. As before, the population mean is $\mu_\emptyset = 50$ and the target values for the subgroup instances c_1, c_2, c_4, and c_8 are $100, 75, 53$, and 12. Additionally, it is assumed that a score of at least 150 is currently required by the result set using the impact interestingness measure q^1_{mean}. For the evaluation of P_A, the instances are added one-by-one, starting with the instance c_1, since it has the highest target value. The interestingness value of a subgroup that covers only this single instance is $q(s_1^{desc}) = 1 \cdot (100 - 50) = 50$. To compute the optimistic estimate $oe^a_{fast}(P_A)$ after the first instance, additionally the value of $oe^a_{remaining}(P_A)$ is required. This is determined as ${i_{P_A}}^a \cdot \left(\frac{\sigma+(i_{P_A}-n)\cdot\theta}{i_{P_1}} - \mu_\emptyset\right) = 4^1 \cdot \left(\frac{100+(4-1)\cdot 100}{4} - 50\right) = 200$. Since 200 exceeds the minimum required interestingness of 150, the evaluation of P_A continues.

Next, the instance c_2 is added, as it has the second highest target value. The corresponding interestingness value is $q(s_2^{desc}) = 2 \cdot (87.5 - 50) = 75$. Additionally, the value of $oe^a_{remaining}(P_A)$ is updated: $oe^a_{remaining}(P_A) = 4^1 \cdot \left(\frac{175+(4-2)\cdot 75}{4} - 50\right) = 125$. It follows that $oe^a_{fast}(P_A) =$

$\max\left(q(s_1^{desc}), q(s_2^{desc}), oe^a_{remaining}(P)\right) = \max(50, 75, 125) = 125$ is an optimistic estimate for the subgroup P_A: P_A itself and all of its specializations are guaranteed to have interestingness scores not higher than 125. As the minimum required interestingness value of the result set is 150, the evaluation of P_A can stop without considering the remaining instances c_4 and c_8.

The formula for the bound described above requires the number of instances covered by a subgroup. This number might not yet be known during the evaluation iteration of the subgroup. However, a simple upper bound for the maximum number of instances in a subgroup can be estimated, e.g., by the known number of instances covered by a generalization of the subgroup.

5.2.4 Optimistic Estimates with Closed Form Expressions

In Section 5.2.2, we discussed a method to derive tight optimistic estimate bounds. However, since these bounds require an ordering of the instances according to the target concept, these bounds cannot be computed with FP-tree-based data representations (Lemmerich et al. 2012). This section presents optimistic estimate bounds that have a closed-form expression that uses only a limited amount of statistics derived from the subgroup. In particular, the bounds for a single subgroup are computable in a distributed single-pass algorithm. Such statistics can also be determined efficiently in FP-tree-based data structures, as shown in previous work (Lemmerich et al. 2012). The information required by the different measures is described in Section 5.1.1.

Mean-based Interestingness Measures

Theorem 10 *As described by Webb (2001), for any subgroup P a tight optimistic estimate for the impact interestingness measure $q^1_{mean}(P) = i_P \cdot (\mu_P - \mu_\emptyset)$ is given by: $oe^1_{mean}(P) = \sum_{c \in P: T(c) > \mu_\emptyset} (T(c) - \mu_\emptyset)$.*

Proof See Webb (2001) for a proof.

The tight optimistic estimate for the binary case presented in (Grosskreutz et al. 2008), i.e., $p_P \cdot (1 - \tau_\emptyset)$, can be seen as special case of this formula, using $T(c) = 1$ for true target values and $T(c) = 0$ for false target values:

$$oe^1_{mean}(P) = \sum_{c \in sg(P), T(c) > \mu_\emptyset} (T(c) - \mu_\emptyset) = \sum_{c \in sg(P), T(c) = 1} (1 - \tau_\emptyset) = p_P \cdot (1 - \tau_\emptyset).$$

This optimistic estimate bound can easily be extended to the other generic mean-based interestingness measures q^a_{mean}:

Theorem 11 *$oe^1_{mean}(P)$ is an optimistic estimate for any generic mean-based interestingness measure $q^a_{mean}(P) = {i_P}^a \cdot (\mu_P - \mu_\emptyset)$ with arbitrary $a \in [0, 1]$.*

Proof For any refinement $r \subseteq sg(P)$ with $q^a_{mean}(r) \geq 0$ and any $a \in [0, 1]$, it holds: $q^a_{mean}(r) = |r|^a(\mu_r - \mu_\emptyset) \leq |r|^1(\mu_r - \mu_\emptyset) = q^1_{mean}(r) \leq oe^1_{mean}(P)$. □

Theorem 12 *An alternative optimistic estimate bound for* $q^a_{mean}(P) = {i_P}^a \cdot (\mu_P - \mu_\emptyset)$ *with arbitrary* $a \in [0,1]$ *is given by:* $\overline{oe}^a_{mean}(P) = {\widetilde{p}_P}^a \cdot (T^{max}_P - \mu_\emptyset)$, *where* $\widetilde{p}_P = |\{c \in sg(P) | T(c) > \mu_\emptyset\}|$ *is the number of instances in the subgroup with a target value higher than the population mean of the target and* T^{max}_P *is the maximum target value in the subgroup.*

Proof It is proven first that no instance with a target value lower than $\mu_\emptyset$ is part of the best refinement: consider any subset $r \subseteq sg(P)$. Then, let $r^+ = \{i \in r | T(c) > \mu_\emptyset\}$ be the set of all instances in r that have a target value higher than the mean of the population and $r^- = \{i \in r | T(c) \leq \mu_\emptyset\}$ the complement of this set, so $r = r^+ \cup r^-$. Then, the interestingness score according to any q^a_{mean} is always equal or higher, if all instances of r^- are removed from the subgroup. So we need to show that:

$$q^a_{mean}(r^+) \geq q^a_{mean}(r)$$
$$|r^+|^a(\mu_{r^+} - \mu_\emptyset) \geq |r|^a(\mu_r - \mu_\emptyset)$$

Similar to the proof of Theorem 10 this can be transformed as follows:

$$|r^+|^a \frac{\sum_{i \in r^+}(T(c) - \mu_\emptyset)}{|r^+|} \geq |r|^a \frac{\sum_{i \in r}(T(c) - \mu_\emptyset)}{|r|}$$
$$|r^+|^a \frac{\sum_{i \in r^+}(T(c) - \mu_\emptyset)}{|r^+|} \geq (|r^+| + |r^-|)^a \frac{\sum_{i \in r^+}(T(c) - \mu_\emptyset) + \sum_{i \in r^-}(T(c) - \mu_\emptyset)}{|r^+| + |r^-|}$$

For shorter notation, we define $S^+ := \sum_{i \in r^+}(T(c) - \mu_\emptyset)$ and $S^- := \sum_{i \in r^-}(T(c) - \mu_\emptyset)$. Due to the construction of r^+ and r^- it holds that $S^+ \geq 0 \geq S^-$. Thus:

$$|r^+|^a \frac{S^+}{|r^+|} \geq (|r^+| + |r^-|)^a \frac{S^+ + S^-}{|r^+| + |r^-|}$$
$$(|r^+| + |r^-|)|r^+|^a S^+ \geq |r^+|(|r^+| + |r^-|)^a (S^+ + S^-)$$
$$(|r^+| + |r^-|)|r^+|^a S^+ \geq |r^+|(|r^+| + |r^-|)^a S^+ + |r^+|(|r^+| + |r^-|)^a S^-$$

Since $S^- \leq 0$, it holds that $((|r^+| + |r^-|)^a S^-) \leq 0$. Thus, the above inequality is satisfied if

$$(|r^+| + |r^-|)|r^+|^a S^+ \geq |r^+|(|r^+| + |r^-|)^a S^+$$
$$(|r^+| + |r^-|)^{1-a} S^+ \geq |r^+|^{1-a} S^+$$
$$(|r^+| + |r^-|)^{1-a} \geq |r^+|^{1-a}$$
$$|r^+| + |r^-| \geq |r^+|$$

This is always true. Therefore, the interestingness value of an instance set r according to any q^a_{mean} never decreases if all instances with a target value less

than the mean target in the overall population are removed. Consequently, there is always a best refinement of a subgroup that does not contain any instance with target value equal or less than the population mean. The largest possible number of instances of such a refinement is given by $\widetilde{p}_P$. Trivially, the mean value of this refinement never exceeds the largest value of the original subgroup. Thus $\widetilde{p}_P \cdot (T_P^{max} - \mu_\emptyset)$ is an optimistic estimate for P. □

Example 5 Once more, we consider the subgroup P_A for the selector sel_A in the toy dataset of Table 1. The population mean in this dataset is $\mu_\emptyset = 50$ and the target values for the subgroup instances c_1, c_2, c_4, and c_8 are $100, 75, 53$, and 12. Then, the optimistic estimate oe^1_{mean} sums over all instances with a target value greater than 50, that is c_1, c_2 and c_4: $oe^1_{mean}(P_A) = (100-50) + (75-50) + (53-50) = 78$. By contrast, the optimistic estimate $\overline{oe}^a_{mean}$ is computed as $\overline{oe}^a_{mean}(P_A) = \widetilde{p}_{P_A}{}^a \cdot (T_{P_A}^{max} - \mu_\emptyset) = 3^a \cdot (100-50) = 3^a \cdot 50$, since three target values are above the population mean. Depending on the generality parameter a of the applied interestingness measures this results in $\overline{oe}^1_{mean}(P_A) = 150$ for $a = 1$, in $\overline{oe}^{0.5}_{mean}(P_A) \approx 86.6$ for $a = 0.5$, or in $\overline{oe}^1_{mean}(P_A) = 50$ for $a = 0$. Comparing these bounds, it is evident that no bound is superior in every case: for a high value of a such as $a = 1$, $oe^1_{mean}(P)$ is tighter than $\overline{oe}^a_{mean}(P_A)$, for a low value of a such as $a = 0$, $\overline{oe}^a_{mean}(P_A)$ is tighter.

As it is evident from the example, both optimistic estimates $oe^1_{mean}(P)$ and $\overline{oe}^a_{mean}(P)$ are not tight for arbitrary parameters a: for the subgroup P_1 the best refinement r^* using the mean test measure $q^{0.5}_{mean}$ contains the first two instances and has the interestingness score $\sqrt{2} \cdot 37.5 \approx 53$, which is lower than both estimates. Both estimates $oe^1_{mean}(P)$ and $\overline{oe}^a_{mean}(P)$ are not exclusive, but can easily be combined: one can compute the values for both bounds and use the tighter one, i.e., the one with the smaller value, to apply optimistic estimate pruning.

Symmetric Mean-based Measures It is easy to extend the optimistic estimates presented above to symmetric variants:

Theorem 13 *An optimistic estimate for the generic symmetric mean-based functions* $q^a_{sym}(P) = i_P{}^a \cdot |\mu_P - \mu_\emptyset|$ *is given by:*
$oe^a_{sym}(P) = \max\left(\sum_{c \in sg(P), T(c) < \mu_\emptyset} (\mu_\emptyset - T(c)), \sum_{c \in sg(P), T(c) > \mu_\emptyset} (T(c) - \mu_\emptyset)\right)$.
This bound is tight for q^1_{sym}. *Another bound is given by:*
$\overline{oe}^a_{sym}(P) = \max\left(\widetilde{p}_P{}^a \cdot (T_P^{max} - \mu_\emptyset), \widetilde{n}_P{}^a \cdot (\mu_\emptyset - T_P^{min})\right)$,
where T_P^{max}, T_P^{min} *are the maximum and minimum target values in the subgroup* P, *and* $\widetilde{p}_P = |\{c \in sg(P) | T(c) > \mu_\emptyset\}|$, $\widetilde{n}_P = |\{c \in sg(P) | T(c) < \mu_\emptyset\}|$ *are the numbers of instances in the subgroup with a target value greater (respectively smaller) than the population mean of target values.*

Proof This follows straightforward from Theorems 11 and 12 since for the interestingness value of any subset $r \subseteq sg(P)$ it holds that $q^a_{sym}(r) = |r|^a |\mu_r -$

$\mu_\emptyset| = max\left(|r|^a(\mu_r - \mu_\emptyset), -(|r|^a(\mu_r - \mu_\emptyset)\right)$. So in essence, one can just compute an upper bound for $q^a_{mean}(r)$ and $-q^a_{mean}(r)$ separately and use the maximum of both bounds as a bound for q^a_{sym}. □

Theorem 14 *For the variance reduction* $q_{vr}(P) = \frac{i_P}{i_\emptyset - i_P} \cdot (\mu_P - \mu_\emptyset)^2$, *two optimistic estimates are given by:*

$$oe_{vr}(P) = \max\left(\frac{i_P}{i_\emptyset - i_P} \cdot (T_P^{max} - \mu_\emptyset)^2, \frac{i_P}{i_\emptyset - i_P} \cdot (T_P^{min} - \mu_\emptyset)^2\right),$$

$$\overline{oe}_{vr}(P) = \max\left(\frac{\left(\sum\limits_{c \in sg(P),\, T(c) > \mu_\emptyset} (T(c) - \mu_\emptyset)\right)^2}{i_\emptyset - 1}, \frac{\left(\sum\limits_{c \in sg(P),\, T(c) < \mu_\emptyset} (T(c) - \mu_\emptyset)\right)^2}{i_\emptyset - 1}\right),$$

where T_P^{max} *(*T_P^{min}*) is the maximum (minimum) target value in the subgroup* P.

Proof The proof of the first estimate is straightforward: The first factor is strictly increasing with i_P and thus reaches its maximum for all specializations of P at i_P, since all specializations cover at most as many instances as P. The maximum difference in the second factor occurs if the mean value in the specialization is either maximal or minimal. Trivially, the maximum or minimum for each specialization is in the interval $[T_P^{min}, T_P^{max}]$.

Regarding the second estimate, it holds for any subset $r \subseteq sg(P)$ with positive interestingness score that: $q_{vr}(r) = \frac{|r|}{i_\emptyset - |r|}(\mu_r - \mu_\emptyset)^2 = \frac{|r|^2}{(i_\emptyset - |r|)|r|}(\mu_r - \mu_\emptyset)^2$ $= \frac{|r|^2}{(i_\emptyset - |r|)|r|}\left(\frac{\sum_{i \in r}(T(c) - \mu_\emptyset)}{|r|}\right)^2 = \frac{1}{(i_\emptyset - |r|)|r|}\left(\sum_{i \in r}(T(c) - \mu_\emptyset)\right)^2$

$(i_\emptyset - |r|)|r|$ gets minimized for $|r| = 1$. The squared sum is maximized if the sum is either maximized or minimized. That is accomplished by either including only positive or only negative summands, that is, only instances with a target value higher than the population mean target value or with a lower target value, respectively. This leads to the presented optimistic estimate. □

Example 6 As in the previous examples, P_A has four instances with target values $T(c_1) = 100$, $T(c_2) = 75$, $T(c_4) = 53$, and $T(c_8) = 12$ in a dataset with an overall target mean of $\mu_\emptyset = 50$. Additionally, it is assumed that the overall population consists of 10 instances. Then, the upper bounds according to the above theorem are given by:

$oe_{vr}(P_A) = \max\left(\frac{4}{8-4} \cdot (100 - 50)^2, \frac{4}{8-4} \cdot (12 - 50)^2\right) = 50^2 = 2500$, and $\overline{oe}_{vr}(P_A) = \max\left(\frac{1}{7} \cdot ((100 - 50)^2 + (75 - 50)^2 + (53 - 50)^2), \frac{1}{7}(12 - 50)^2\right) = \frac{1}{7} \cdot (50^2 + 25^2 + 3^2) \approx 447.7$. In this case, the second bound is substantially tighter (lower) and is therefore used for pruning.

Median-based Measures For median-based interestingness measures, the practical use of a direct estimate that can be computed in a parallel single pass algorithm is doubtful since the median itself cannot be computed in such a way. Nonetheless, a very simple, but loose estimate can be specified:

Theorem 15 *An optimistic estimate for the generic median-based measure* $q^a_{med}(P) = {i_P}^a \cdot (med_P - med_\emptyset)$ *is given by:* $oe^a_{med}(P) = {i_P}^a \cdot (T^{max}_P - med_\emptyset)$.

Proof The maximum median in any refinement cannot exceed the maximum occurring value in the subgroup, and the size of a refinement cannot exceed the size of the subgroup. □

In contrast to the generic mean-based functions, the best refinement for median-based function can contain values with target values lower than the population mean as demonstrated in the following example.

Example 7 Consider a subgroup P_2 with target values $\{3, 2, 2, 0, -1, -1\}$ in a dataset with an overall median target value of 1. Then, for the interestingness measure q^1_{med} the best subset of instances contains the first 5 instances and has an interestingness score of 5.

(Full) Distribution-based Measures

Theorem 16 *An optimistic estimate for the Kolmogorov-Smirnov interestingness measure* $q_{ks}(P) = \sqrt{\frac{i_P \cdot i_{\neg P}}{i_\emptyset}} \Delta_{(P,\neg P)}$ *for any subgroup* P *with* $i_P < \frac{i_\emptyset}{2}$ *is given by* $oe_{ks}(P) = \sqrt{\frac{i_P(i_\emptyset - i_P)}{i_\emptyset}}$.

Proof The interestingness value of q_{ks} is given by $\sqrt{\frac{i_P \cdot i_{\neg P}}{i_\emptyset}} \cdot \Delta_{(P,\neg P)}$. The test statistic $\Delta_{(P,\neg P)}$ is computed as the supremum of differences in the empirical distribution functions of P and its complement. Since the range of the empirical distribution function is $[0, 1]$, the supremum of the difference $\Delta_{(P,\neg P)} \leq 1$. For a fixed population, the left term $\sqrt{\frac{i_P \cdot i_{\neg P}}{i_\emptyset}}$ is only dependent on the number of instances covered by the subgroup. We determine the maximum of this term for any refinement r of $sg(P)$. If $i_P \leq \frac{i_\emptyset}{2}$ the term is monotone. In particular, $|r|(i_\emptyset - |r|) < i_P(i_\emptyset - i_P)$. Otherwise a maximum is reached at $i_P = \frac{i_\emptyset}{2}$. However, this is an overall bound for the interestingness measure and is therefore not useful for pruning. □

Given a minimum interestingness value required by the result set, the optimistic estimate derived from this theorem implies a minimum number of instances that must (at least) be covered by any subgroup that has a sufficiently high interestingness score. In contrast to the other introduced optimistic estimates, this bound does not take the distribution of the target variable in the subgroup into account. Therefore, it could be expected that this bound is less tight than other optimistic estimates.

Example 8 Consider again the subgroup P_A that covers 4 instances of the 8 instances in the overall dataset. The optimistic estimate for the Kolmogorov-Smirnov interestingness measure is then given by: $\sqrt{\frac{4 \cdot (8-4)}{8}} = \sqrt{2}$. The target values of the instances covered by the subgroup do not influence the optimistic estimate.

Rank-based Measures

Theorem 17 *Two optimistic estimate bounds for the Mann-Whitney interestingness measure* $q_{mw'}(P) = \sqrt{\frac{i_P}{i_{\neg P}}} \cdot (\frac{\mathcal{R}_P}{i_P} - \frac{i_\emptyset+1}{2})$ *are given by:*

$$oe^1_{mw'}(P) = \sum_{c \in sg(P), \rho(c) > \frac{i_\emptyset+1}{2}} \left(\rho(c) - \frac{i_\emptyset + 1}{2} \right)$$

$$\overline{oe}_{mw'}(P) = \sqrt{i_P^+}(\rho_P^{max} - \frac{i_\emptyset + 1}{2}),$$

where $\rho(c)$ *is the rank of instance* c *in order of the target values,* ρ_P^{max} *is the maximum rank in the subgroup* P *and* $i_P^+ = |\{c \in sg(P) \,|\, \rho(c) > \frac{i_\emptyset+1}{2}\}|$ *is the number of instances in the subgroup with a rank higher than the population's rank mean.*

Proof It holds for all refinements r with a positive interestingness value, that

$$q_{mw'}(P) = \sqrt{\frac{i_P}{i_{\neg P}}}(\frac{\mathcal{R}}{i_P} - \frac{i_\emptyset + 1}{2}) \leq \sqrt{i_P}(\frac{\mathcal{R}}{i_P} - \frac{i_\emptyset + 1}{2}).$$

Since $\frac{\mathcal{R}}{i_P}$ is the mean of the ranks within the subgroup and $\frac{i_\emptyset+1}{2}$ is the mean of the ranks in the overall population, the right part of this equation is equal to the mean test function $q^{0.5}_{mean}$ if the target values are given by the ranks. Thus, we can transfer the upper bounds from Theorem 11. However, these bounds are substantially less tight for this interestingness measure due to the initial estimation. □

Example 9 Consider again the subgroup P_A in a dataset of 8 instances and assume that the 4 instances covered by the subgroup have the ranks (in ascending order of target values) $\rho(c_1) = 8$, $\rho(c_2) = 7$, $\rho(c_4) = 5$ and $\rho(c_8) = 1$. That means, for example, that the instance c_2 has the seventh lowest target value in the dataset. Then, the optimistic estimates according to the above theorem are given by:

$$oe^1_{mw'}(P_A) = (8 - \frac{9}{2}) + (7 - \frac{9}{2}) + (5 - \frac{9}{2}) = 6.5$$

$$\overline{oe}_{mw'}(P_A) = \sqrt{3} \cdot (8 - \frac{9}{2}) = \sqrt{3} \cdot 3.5$$

The second bound is tighter in this example and is therefore used as the overall bound for optimistic estimate pruning.

Theorem 18 *If there are no ties in the ranks, then an optimistic estimate for the area-under-the-curve interestingness measure* $q_{auc}(P) = \dfrac{\mathcal{R}_{\neg P} - \frac{i_{\neg P} \cdot (i_{\neg P}+1)}{2}}{i_P \cdot i_{\neg P}}$ *is given by:* $oe_{auc}(P) = \frac{i_\emptyset - \rho_P^{min}}{i_\emptyset - 1}$, *where* ρ_P^{min} *is the minimum rank for an instance in* P.

Proof Due to the construction of the area-under-the-curve the best subset of $sg(P)$ contains only one instance, which is the one with the lowest rank ρ_{min}. The interestingness of this refinement S is $q_{auc}(S) = \frac{\mathcal{R}_{\neg S} - \frac{i_{\neg S}(i_{\neg S}+1)}{2}}{i_S i_{\neg S}} = \frac{\frac{(i_\emptyset+1)i_\emptyset}{2} - \rho_P^{min} - \frac{i_\emptyset(i_\emptyset-1)}{2}}{1 \cdot (i_\emptyset - 1)} = \frac{i_\emptyset - \rho_P^{min}}{i_\emptyset - 1}$ □

Note that these bounds in closed form for the measure q_{auc} are not guaranteed to return optimal results in case of ties.

Example 10 Consider again the subgroup P_A that covers instances with following ranks with respect to the target concept. $\rho(c_1) = 8$, $\rho(c_2) = 7$, $\rho(c_4) = 5$ and $\rho(c_8) = 1$ in a dataset that consists of 8 instances. Then, the above theorem provides the optimistic estimate: $oe_{auc}(P) = \frac{8-1}{7} = 1$. Since this is an overall bound for the interestingness measure, pruning cannot be applied here.

5.3 Algorithms for Subgroup Discovery with Numerical Targets

Below, we present two novel algorithms for efficient subgroup discovery that integrate the presented approaches regarding data structures and optimistic estimate bounds. Both algorithms employ the same enumeration strategy, that is, depth-first-search with one level look-ahead, but use different data structures and – as a consequence – different optimistic estimate bounds.

5.3.1 The SD-Map* *Algorithm*

The *SD-Map** algorithm improves its predecessor SD-Map (Atzmueller and Puppe 2006) in several directions: while SD-Map focuses exclusively on binary targets, *SD-Map** extends the employed FP-tree data structure in order to determine statistics of the numerical target concept as described in Section 5.1.1. The statistics contained in the nodes of the FP-trees are used to compute not only the interestingness of subgroups, but also their optimistic estimates. In that direction, *SD-Map** allows the incorporation of pruning based on the bounds in closed-form expressions, which have been presented in Section 5.2.4. Ordering-based bounds cannot be applied since the ordering information is not captured by FP-tree representations, see (Lemmerich et al. 2012) for a formal proof.

Pruning is applied in two different forms within the algorithm: first, *selector pruning* is performed in the recursive step, when a conditional FP-tree is built. A (conditioned) branch is omitted if the optimistic estimate for the conditioning selector is below the threshold given by the k best subgroup qualities. Second, *header pruning* is used, when a (conditional) frequent pattern tree is constructed. Here, all the nodes with an optimistic estimate below the mentioned interestingness threshold can be omitted.

To maximize the efficiency of pruning, the search strategy was also slightly modified: instead of the basic depth-first-search used in the SD-Map algorithm,

*SD-Map** applies a modified depth-first strategy with look-ahead, similar to the *DpSubgroup* algorithm in (Grosskreutz et al. 2008). Reordering of the search space is performed by sorting of the header nodes: during the iteration over the candidate selectors for the recursive call, the selectors are reordered according to their optimistic estimate value. In doing so, more promising selectors are evaluated first. In a top-k approach, this helps to include high scoring subgroups early into the result set in order to provide higher interestingness thresholds for more efficient pruning.

5.3.2 The NumBSD *Algorithm*

Although FP-tree-based approaches have shown excellent performance, they are not applicable to all interestingness measures. Additionally, they cannot make use of the tighter ordering-based pruning schemes. Furthermore, the construction of an initial FP-tree can require significant overhead, particularly if the search is limited to small search depths. Therefore, we present the exhaustive subgroup discovery algorithm *NumBSD* as an alternative: it uses an efficient vertical, bitset-based data structure as described in the previous section. As search strategy, *NumBSD* employs a depth-first-search approach with one level look-ahead, similar to the *SD-Map** algorithm. The algorithm applies efficient pruning strategies, including ordering-based bounds and fast bounds, see Sections 5.2.2 and 5.2.3.

The algorithm *NumBSD* and its sub-procedures are shown in Algorithm 1. It first initializes the vertical data structures and then calls the main recursive function *recurse*. This function consists of two parts. In the first part (lines 2-9) all direct specializations, that is, all subgroups created by adding a single selector to the description of the current subgroup, are considered. For these specializations the corresponding bitsets, the interestingness value and the optimistic estimates are computed. This is achieved efficiently in a single run through the subgroup by the method *computeRefinement* described below. If the interestingness value of a specialization is sufficiently high, then it is added to the result set. This potentially replaces a subgroup with a lower interestingness score and increases the minimum required interestingness score of the result set. Only if the optimistic estimate for a specialization exceeds the minimum interestingness value of the result set, then this refinement is also considered for the recursive search. In the second part of the function (lines 10-13), it calls itself recursively for these candidate subgroups.

For the performance of the algorithm, an efficient computation of the bitset, the interestingness score and the optimistic estimate of a refinement is essential. This is performed in the function *computeRefinement* of Algorithm 1. First, an upper bound for the maximum number of instances of a refinement is given by the number of instances for the current subgroup and for the additional selector. Each bitset technically consists of words of 32 (respectively 64) bits. The bitset representing the instances of the specialization *spec* is computed word by word by a logical *AND* between the bitset of the current subgroup and the bitset of the new selector. Then, for each bit in this word

Algorithm 1 *NumBSD* algorithm

```
 1: function NumBSD(maxDepth)
 2:     Sort(allInstances) // sort descendingly w.r.t. target values
 3:     TargetVal ← array of target values
 4:     for all sel in allSelectors do
 5:         bitsets(sel) ← createBitset(sel)
 6:     allTrue ← new bitset, all bits set to 1
 7:     recurse(allTrue, ∅, allSelectors, maxDepth)

 1: function recurse(currentBitset, currentDescription, remainingSels, maxDepth)
 2:     nextSelectors ← ∅
 3:     for all sel in remainingSels do
 4:         nextBitSet ← computeRefinement(currentSG,sel, result.minQ)
 5:         if nextBitSet.estimate > result.minQ then
 6:             nextBitsets(sel) ← nextBitset
 7:             nextSelectors ← nextSelectors ∪ sel
 8:             if nextBitSet.quality > result.minQ then
 9:                 result.add (currentDescription ∪ sel)
10:     if prefix.size < maxDepth then
11:         Sort(nextSelectors) // w.r.t. optimistic estimates
12:         for all sel in nextSelectors do
13:             recurse(nextBitsets(sel), prefix ∪ sel, nextSelectors \ sel, maxDepth)

 1: function computeRefinement(currentBitset, sel, minQualThreshold)
 2:     maxN ← Math.min(currentBitset, bitsets(sel).cardinality)
 3:     n ← 0
 4:     sum ← 0
 5:     maxEstimate ← 0;
 6:     refinement ← new bitset ()
 7:     for all i = 0 to countWords (currentBitset) do
 8:         refinement.word[i] ← currentBitset.word[i] AND bitsets(sel).word[i]
 9:         for all each bit b in refinement.bitset.word[i], that is set to true do
10:             n ← n+1
11:             currentValue ← TargetVal [global position of b]
12:             sum ← sum + currentValue
13:             maxEstimate = max (maxEstimate, computeQuality (n, sum))
14:         sumEstimateAtEnd = sum + currentValue · (maxN - n);
15:         maxOEatEnd = computeQuality (maxN, sumEstimateAtEnd)
16:         if (maxEstimate < minQ ∧ maxOEatEnd < minQ) then
17:             refinement.optEstimate = max (maxEstimate, maxOEatEnd)
18:             return refinement % exploit fast pruning bounds
19:     return refinement
```

that is set to true (each instance of the refinement) the count and sum of target values are adjusted. Based on these values the interestingness score of the current part of the refinement is computed. When considering the i-th bit of the refinement, the interestingness score is equal to $q(s_i^{desc})$ for the refinement *spec* in the terminology of Theorem 1. Thus the maximum of the interestingness scores computed in this way determines a tight optimistic estimate for the subgroup under evaluation. Since $s_{i_P}^{desc} = sg(P)$, the last of these scores is equal to the interestingness of the overall subgroup.

As a further improvement, the fast optimistic estimates, cf. Theorem 9 are checked after each word of the bitset. If neither this bound nor any of the al-

Table 2 A summary of interestingness measure with respect to properties regarding efficient computation.

Measure	Notation	Formula	Estimate in closed form	Estimable by ordering	Computable in *SD-Map**
Impact	$q^1_{mean}(P)$	$i_P(\mu_P - \mu_\emptyset)$	yes	one-pass	yes
Mean-based	$q^a_{mean}(P)$	${i_P}^a(\mu_P - \mu_\emptyset)$	yes	one-pass	yes
z-score	$q_z(P)$	$\sqrt{i_P}\frac{(\mu_P - \mu_\emptyset)}{\sigma_0}$	yes	one-pass	yes
Variance-based	$q^a_\sigma(P)$	${i_P}^a(\sigma_P - \sigma_\emptyset)$	no	no	yes
t-score	$q_t(P)$	$\sqrt{i_P}\frac{(\mu_P - \mu_\emptyset)}{\sigma_P}$	no	no	yes
Sym. mean-based	$q^a_{sym}(P)$	${i_P}^a\lvert\mu_P - \mu_\emptyset\rvert$	yes	two-pass	yes
Variance reduction	$q_{vr}(P)$	$\frac{i_P}{(i_\emptyset - i_P)}(\mu_P - \mu_\emptyset)^2$	yes	two-pass	yes
Generic median	$q^a_{med}(P)$	${i_P}^a(med_s - med_0)$	(yes)[1]	one-pass	no
Kolmogorov-Smirnov	$q_{ks}(P)$	$\sqrt{\frac{i_P \cdot i_{\neg P}}{i_\emptyset}}\Delta_{P,\neg P}$	yes	no[2]	no
Mann-Whitney	$q_{mw}(P)$	$\sqrt{\frac{i_P}{i_{\neg P}}}(\frac{\mathcal{R}}{i_P} - \frac{i_\emptyset}{2})$	yes	one-pass	yes
AUC	$q_{auc}(P)$	$\frac{\bar{\mathcal{R}} - \frac{i_{\neg P} i_{\neg P}+1}{2}}{i_P i_{\neg P}}$	yes	one-pass	yes

ready computed values $q(s_i^{desc})$ are sufficiently high for the result set, then the evaluation of the current subgroup can stop. Since the subgroup itself and all of its generalization will not contribute to the result set, the exact values of the interestingness and the optimistic estimate are not of interest anymore. Thus, parts of this computation can be safely omitted using the fast bounds introduced previously. In the worst case, the method *computeRefinement* requires one complete pass through the instances of the current subgroup.

This paper focuses on bitsets as vertical data representations. However, a very similar algorithm could be obtained by employing other vertical data structures such as TID-lists, see Zaki (2000). These could be expected to perform faster if the data is sparse, i.e., if selectors cover only small fractions of the dataset.

5.4 Summary: Interestingness Measures and their Computational Properties

In the previous sections, we showed that efficiency optimizations for subgroup discovery with numerical target concepts do strongly depend on the applied interestingness measures. Some interestingness measures, such as the t-score, can be determined by *SD-Map**, taking full advantage of the more sophisticated, compressed FP-tree data structure. Other interestingness measures in

[1] The generic measure itself cannot be computed by *SD-Map**.

[2] Not yet determined.

turn, cannot be determined by *SD-Map** at all, e.g., median-based measures. Additionally, ordering-based optimistic estimate bounds could be derived for a large variety of interestingness measures, but for a few exceptions this was not possible, e.g., for the t-score.

Table 2 summarizes interestingness measures with respect to their computational properties. In particular, the table shows for each interestingness measure if there is an optimistic estimate in closed form presented in this work that can be computed using FP-trees, if it is estimable by ordering, and if the measure itself is computable by the *SD-Map** algorithm.

6 Evaluation

The benefits of the proposed improvements were evaluated in a wide range of experiments. The algorithms were implemented in the open source subgroup discovery environment *VIKAMINE*[3], see (Atzmueller and Lemmerich 2012). Runtime experiments were performed on a standard office PC with a 2.2 GHz CPU and 2 GB RAM. Experiments that count the number of evaluated candidates were executed on additional machines since results are independent of the hardware performance.

The experiments used publicly available datasets from the UCI (Lichman 2013) and KEEL (Alcala-Fernandez et al. 2011) data repositories. For nominal attributes, attribute-value pairs were used as selectors. Numerical attributes in the search space were discretized into ten intervals by equal-frequency discretization. Runtimes are reported for the full algorithms including the initial sorting step (if required), but excluding loading and pre-processing such as determining the selector set for the search. No overlapping intervals were generated. Due to their popularity, a focus of the experiments is on generic mean-based interestingness measures, cf. (Klösgen 1996; Wrobel 1997; Webb 2001; Grosskreutz 2008; Atzmueller and Lemmerich 2009; Lemmerich and Atzmueller 2012; Atzmueller and Lemmerich 2013).

The evaluation section is structured as follows: we first investigate the influences of the adapted data structures and optimistic estimate bounds separately, before the runtimes of the full-featured algorithms are compared and specific findings on *SD-Map** and *NumBSD* are discussed. Then, the influence of the result set size and the impact of the bounds using limited information, see Section 5.2.3, are evaluated. Finally, we summarize the experimental results, and discuss implications for the application of the proposed algorithms.

6.1 Effects of Optimistic Estimates

The first set of experiments evaluated the use of the introduced optimistic estimate bounds. In that direction, a subgroup discovery algorithm with depth-first-search with one level look-ahead and no reordering of the search space,

[3] Available at *www.vikamine.org*

cf. Section 5.3.1, was run. Several interestingness measures were tested with a fixed maximum search depth, that is, a maximum numbers of selectors in a description, of $d = 5$ ($d = 4$ for two datasets due to long runtimes). For the optimistic estimate pruning, a top-1 approach was applied, that is, only the one subgroup with the top score was sought. Each run was executed three times in different variations: with no optimistic estimate pruning, with optimistic estimate pruning using the bounds in closed form and with ordering-based optimistic estimate pruning. For each variation, the number of evaluated candidates in recursive calls, that is, after the initial evaluation of the basic selectors (not including these), was counted. The respective results are summarized in Table 3, and Table 4.

It is evident that the number of required evaluations is reduced substantially by applying the presented optimistic estimate bounds. Regarding different mean-based interestingness measures, optimistic estimate pruning has generally less impact if the parameter a in the interestingness measures is lower (e.g., $a = 0.5$ and $a = 0.1$), that is, if deviations of the target concept are more important (see Table 3). This can be explained by the fact that even small subgroups can achieve high scores in this scenario and thus the anti-monotonicity of the subgroup size is more difficult to exploit in these cases. An exception is the extreme parameter $a = 0$, which is equivalent to the *average interestingness measure*: since the best refinement of a subgroup is already determined by the single instance with the highest target value, all subgroups that do not cover one of the instances with high target values can immediately be pruned by applying the novel estimates of Theorem 12.

For the extreme settings $a = 1$ and $a = 0$, the bounds in closed form are tight, that is, they allow for the same amount of pruning as ordering-based bounds. For the intermediate settings $a = 0.5$ and $a = 0.1$, bounds in closed forms are considerably less precise. Therefore, often substantially more candidates must be evaluated in comparison to ordering-based bounds. However, the optimistic estimate bounds in closed form still reduce the number of required evaluations by orders of magnitude in comparison to the unpruned search space. Note that ordering-based bounds cannot be combined with all data structures and come at higher computational costs.

Also for other interestingness measures, see Table 4, applying optimistic estimate bounds can lead to a significant reduction of necessary subgroup evaluations. However, the amount of that reduction is of course heavily influenced by the utilized interestingness measure. For the symmetric mean-based measure and the variance reduction, the number of evaluated candidates is often decreased to less than 1000, if ordering-based optimistic estimates are applied. The optimistic estimates in closed form are less tight, but still reduce the number of required evaluations by an order of magnitude or more. Ordering-based bounds are also very effective for the other investigated interestingness measures, that is, the median-based measure $q^{0.5}_{med}(P)$, the Mann-Whitney measure $q_{mw}(P)$, and the area-under-the-curve $q_{auc}(P)$. Regarding optimistic estimates in closed form, even relatively simple-to-derive bounds can reduce the number of required candidate evaluations substantially, as indicated by the results for

Table 3 Comparison of pruning schemes, i.e., no pruning (None), ordering-based bounds (Order.) and bounds in closed form (Closed). The table provides numbers of subgroups that had to be evaluated in a depth-first-search with one level look-ahead and no reordering of the search space (not counting the basic selectors). The search was restricted to a maximum of 5 selectors (only 4 for the two datasets marked with a "*" due to long runtimes). For the applied mean-based interestingness measures, different parameters a were evaluated as indicated in the column headers. If no pruning is applied, the number of required candidate evaluations is independent from the applied interestingness measure.

Dataset		1.0		0.5		0.1		0.0	
	None	Closed	Order.	Closed	Order.	Closed	Order.	Closed	Order.
adults	8,503,218	253	253	87,800	1,872	410,321	32,873	1,252	1,252
ailerons	984,289,405	4,298	4,298	912,256	4,384	46,081,383	148,830	1,470	1,470
autos	12,316,190	17	17	8,347	66	52,764	1,884	134	134
breast-w	219,993	12	12	427	19	6,413	963	132	132
census-kdd*	73,374,193	1,815	1,815	262,179	19,012	3,576,514	145,315	1,554	1,554
communities*	$> 2 \cdot 10^9$	22	22	384,142	92	22,581,558	79,187	2,248	2,248
concrete_data	209,041	19	19	2,532	131	3,936	707	457	457
credit-a	2,231,118	277	277	26,358	940	17,105	867	777	777
credit-g	8,389,271	364	364	88,883	906	84,227	383	319	319
diabetes	350,466	17	17	5,113	102	2,008	316	342	342
elevators	121,983,859	62	62	54,677	388	214,793	2,389	1,133	1,133
flare	101,946	29	29	446	446	1,497	1,497	5	5
forestfires	1,209,242	30	30	1,090	172	875	164	161	161
glass	141,714	13	10	81	17	529	87	147	147
heart-c	823,995	224	224	14,086	975	5,942	516	357	357
house	173,768,450	21	21	182,583	482	465,228	13,424	769	769
housing	1,554,972	43	43	4,885	387	4,011	273	127	129
letter	69,157,431	8	8	10,226	9	183,379	3,409	988	990
mv	5,542,943	34	29	60	35	2,262	34	1,499	1,499
pole	47,553,142	48,077	48,077	144,353	72,851	650,349	146,175	270	270
sonar	1,737,064,885	12	12	15,521	12	156,647	1,253	1,208	1,208
spambase	1,045,755,337	2,741,042	2,741,042	44,663,301	6,726,031	11,381,117	2,923,091	878	878
ticdata	1,254,395,632	1,783,651	1,783,651	249,736,031	5,658,523	288,205,279	6,370,265	61	61
yeast	291,042	23	23	1,941	27	3,190	259	217	217

Table 4 Effects of the introduced optimistic estimate bounds for further interestingness measures, i.e., a symmetric mean-based measure ($q^{0.5}_{sym}$), the variance reduction (q_{vr}), a median-based measure ($q^{0.5}_{med}$), the Kolmogorov-Smirnov measure (q_{ks}), the Mann-Whitney (q_{mw}) measure, and the Area-under-the-Curve measure (q_{auc}). Depending on the interestingness measure, ordering-based bounds (Order.), bounds in closed form (Closed), or both were tested. The table provides numbers of subgroups that had to be evaluated in a depth-first-search with one level look-ahead and no reordering of the search space. The search was restricted to a max. of 5 selectors (only 4 for the datasets marked with a "*" due to long runtimes). Bounds in closed form for the measure q_{auc} are not guaranteed to return optimal results in case of ties.

Dataset	None	$q^{0.5}_{sym}$ Closed	$q^{0.5}_{sym}$ Order.	q_{vr} Closed	q_{vr} Order.	q_{iv} Order.	q_{med} Order.	q_{ks} Closed	q_{mw} Closed	q_{mw} Order.	q_{auc} Closed	q_{auc} Order.
adults	8,503,218	114,884	313	174,961	195	195	16,962	78,546	6,086,727	17,900	8654	339
ailerons	984,289,405	11,580,481	49,312	16,419,078	13,905	13,905	13,974	2,263,350	949,909,065	28,175	12049	408
autos	12,316,190	24,917	69	29,252	66	66	4,089	68,856	4,994,035	4,769	1651	347
breast-w	219,993	8,985	14	9,801	17	17	4,094	15,541	180,065	4,418	995	896
census.*	73,374,193	1,488,608	34,599	1,846,126	3,977	3,977	19,281	86,246	63,134,807	15,306	106	856
comm.*	$> 2 \cdot 10^9$	4,184,249	94	13,070,933	110	110	27,150	6,501,639	$> 2 \cdot 10^9$	47,376	5788	638
concrete.	209,041	13,165	28	15,625	21	21	4,173	37,009	115,277	5,385	4438	412
credit-a	2,231,118	87,980	1,054	120,844	885	885	16,663	210,403	1,237,561	23,015	8812	276
credit-g	8,389,271	297,275	906	446,215	867	867	20,186	1,175,479	5,489,826	46,207	2760	310
diabetes	350,466	15,451	117	19,494	84	84	8,961	78,174	177,820	11,241	1330	486
elevators	121,983,859	311,094	388	886,706	290	290	10,173	1,042,651	78,737,713	30,138	3636	716
flare	101,946	3,062	446	7,322	380	380	2,796	55,649	21,208	2,544	270	1364
forestfires	1,209,242	2,492	171	2,492	168	168	1,254	180,998	570,104	9,454	365	637
glass	141,714	3,743	29	6,006	25	25	1,613	23,831	60,032	3,165	451	993
heart-c	823,995	33,453	658	41,367	654	654	8,070	107,226	475,716	10,594	2900	264
house	173,768,450	733,605	509	1,041,836	401	401	40,540	196,344	111,649,531	19,420	1977	735
housing	1,554,972	10,339	382	10,831	226	226	4,862	25,462	438,041	3,430	1543	380
letter	69,157,431	93,471	15	150,737	7	7	32,817	144,219	44,679,902	26,358	2225	613
mv	5,542,943	69,771	43	120,249	41	41	7,520	76,108	4,053,866	8,238	24385	492
pole	47,553,142	190,154	72,850	1,300,541	70,712	70,712	163,186	1,477,117	27,444,125	300,872	308	2664
sonar	1,737,064,885	185,796	5	241,948	5	5	11,107	1,885,904	657,387,130	75,905	3279	552
spambase	1,045,755,337	45,509,446	6,726,031	76,577,712	6,709,191	6,709,191	5,456,553	5,288,584	687,069,672	7,088,977	796	2556
ticdata	1,254,395,632	220,568,015	6,659,758	253,256,371	5,528,202	5,528,202	3,715,310	4,218,711	961,636,828	4,662,825	19873	504
yeast	291,042	17,578	28	21,942	28	28	2,868	54,281	168,478	4,197	3190	439

the Kolmogorov-Smirnov interestingness measure. The least effective bounds were by far the optimistic estimates for the Mann-Whitney, which only was able to prune about 40% of the candidates on average. Overall, the reduction of required candidate evaluations was substantial for almost all evaluated interestingness measures and datasets. The remainder of the evaluation will focus on the mean-based interestingness measures with different parameterizations.

6.2 Influences of Data Structures

In the next series of experiments, the effects of different data structures were investigated. Regarding that aspect, the runtimes of the presented algorithms *without* applying optimistic estimate pruning were measured. This was performed for the *NumBSD* algorithm, which is based on a bitset-based representation, as well as for the *SD-Map** algorithm, which is based on FP-trees. For comparison, the task was also solved by a simple depth-first-search without any specialized data structure (repeated checking of the selection expressions in memory). Since no optimistic estimate bounds are exploited, the runtime was (almost) independent from the applied interestingness measure and size of the result set k, cf. also (Lemmerich et al. 2012). The experiments were performed with different maximum search depths $d = 2, \ldots, 6$. Table 5 displays representative results for the measure $q^{0.5}_{mean}$ and a result set size of $k = 1$.

The results show that both introduced data structures – the bitset-based structure as well as the FP-tree-based representation – substantially outperform the simple approach. A direct comparison between the two approaches is more difficult: for lower search depths ($d = 2, 3, 4$), bitset-based structures usually enable faster runtimes than FP-trees. The differences reach an order of magnitude for some datasets, e.g., for the *communities* and *spambase* datasets. For higher search depths ($d = 5, 6$), the results are more ambiguous: for some datasets the bitsets perform better, for some they perform worse than FP-trees. In particular, for datasets with a high instance count, the FP-tree-based approach is able to finish the tasks fast. In the *census-kdd* dataset, which is the largest tested dataset (in terms of instances), FP-trees perform better than bitsets already at a search depth of 4. This can be explained by the fact that the FP-trees achieve a better compression of the data in datasets with a high instance count. Additionally, for higher search depths, subgroups cover only small parts of a dataset leading to sparsely populated bitsets. In these cases, using TID-lists (cf. Zaki (2000)) instead of bitsets might lead to improved runtimes. This is to be explored in future work.

In summary, FP-trees are the data structure of choice if the dataset contains many instances, and if the maximum allowed number of selectors in a description is large. By contrast, bitsets are preferred if the search is restricted to low search depths, or if the instance count is comparatively low. For some interestingness measure, it is not possible to derive optimistic estimate bounds, e.g., for generic variance-based measures or the t-score. Therefore, the run-

Table 5 Comparison of data structures: the table shows the runtimes in seconds of simple depth-first-search (Simple), *NumBSD* (BSD), and *SD-Map** (SDM) *without optimistic estimate pruning* for different maximum search depths d. Here, results are shown for the interestingness measure $q^{0.5}_{mean}$, but results are very similar for all applicable interestingness measures.

Dataset	instances	$d=2$			$d=3$			$d=4$		$d=5$		$d=6$	
		Simple	BSD	SDM	Simple	BSD	SDM	BSD	SDM	BSD	SDM	BSD	SDM
adults	32561	98.0	1.5	4.3	3068.8	4.5	7.5	24.6	16.3	103.8	42.5	300.5	86.3
ailerons	13750	394.7	1.5	47.5	>4 h	38.4	268.5	1009.0	1748.6	>4 h	10554.0	>4 h	>4 h
autos	205	1.3	< 0.1	0.3	41.4	0.6	1.7	4.3	7.5	20.6	36.3	75.6	145.0
breast-w	699	0.5	< 0.1	0.1	7.5	0.1	0.1	0.2	0.2	0.3	0.4	0.5	0.5
census-kdd	199523	8465.7	47.2	223.4	>4 h	533.2	1039.1	10470.0	5436.0	>4 h	>4 h	>4 h	>4 h
communities	1994	332.7	3.3	482.3	>4 h	567.3	10505.1	>4 h	>4 h	>4 h	>4 h	>4 h	>4 h
concrete_data	1030	0.6	< 0.1	0.1	9.7	0.1	0.2	0.2	0.3	0.4	0.5	0.4	0.5
credit-a	690	0.8	< 0.1	0.2	16.4	0.2	0.7	1.0	2.2	4.1	6.6	10.8	14.8
credit-g	1000	1.4	< 0.1	0.7	36.2	0.3	3.0	3.1	11.6	18.8	40.8	77.5	128.8
diabetes	768	0.4	< 0.1	0.1	7.4	0.1	0.2	0.3	0.4	0.5	0.5	0.6	0.6
elevators	16599	49.2	0.5	9.7	2198.8	5.6	28.9	83.8	96.4	672.1	309.0	2627.7	887.0
flare	1066	0.4	< 0.1	0.1	3.2	< 0.1	0.1	0.1	0.2	0.3	0.4	0.7	0.8
forestfires	517	1.1	< 0.1	0.2	27.4	0.2	0.5	0.9	1.3	2.0	2.9	3.4	4.4
glass	214	0.2	< 0.1	< 0.1	2.5	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3
heart-c	303	0.2	< 0.1	0.1	3.7	0.1	0.2	0.4	0.7	1.3	1.8	2.7	3.4
house	22784	70.1	0.8	16.9	3077.3	7.7	51.9	131.6	157.2	1149.7	464.3	3565.7	1060.1
housing	506	0.8	< 0.1	0.2	19.3	0.2	0.5	1.0	1.4	2.5	3.5	4.6	6.3
letter	20000	54.1	0.7	10.6	2157.8	5.3	31.5	64.0	92.5	436.6	275.2	1455.9	601.3
mv	40768	31.4	0.8	5.1	605.7	2.2	8.2	13.5	14.3	76.0	27.5	196.7	44.6
pole	14998	25.0	0.5	7.1	762.1	4.2	27.2	47.4	112.6	430.7	447.8	2894.4	1725.9
sonar	208	9.5	0.5	8.5	1480.1	28.4	100.9	353.1	780.2	2971.2	9084.8	>4 h	>4 h
spambase	4601	25.4	0.5	30.6	1650.4	11.9	287.0	266.8	2521.1	4626.0	>4 h	>4 h	>4 h
ticdata	5822	244.6	2.0	155.2	>4 h	91.0	2018.7	3414.8	>4 h	>4 h	>4 h	>4 h	>4 h
yeast	1484	0.8	< 0.1	0.1	11.3	0.1	0.2	0.3	0.4	0.5	0.7	0.7	0.7

Table 6 Comparison of the full algorithms: the table shows the runtimes in seconds for *NumBSD* (BSD) and for *SD-Map** (SDM) with *all pruning options enabled* for different interestingness measures. The search was limited to a maximum search depth of $d = 5$. The first two columns show results of the algorithms without optimistic estimate pruning for comparison.

Dataset	No Pruning		q^1_{mean}		$q^{0.5}_{mean}$		$q^{0.1}_{mean}$		q^0_{mean}	
	BSD	SDM	BSD	SDM	BSD	SDM	BSD	SDM	BSD	SDM
adults	103.8	42.5	0.8	2.3	11.1	6.1	19.7	14.1	1.3	2.5
ailerons	>4 h	10554.0	6.0	1.7	45.3	17.7	95.5	3345.9	0.6	147.0
autos	20.6	36.3	< 0.1	< 0.1	< 0.1	0.1	< 0.1	0.3	< 0.1	< 0.1
breast-w	0.3	0.4	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
census-kdd	>4 h	>4 h	53.3	52.6	5184.3	1110.2	> 4 h	5795.2	35.0	52.1
communities	>4 h	>4 h	0.1	0.9	0.1	4.3	2.0	108.0	0.3	0.9
concrete_data	0.4	0.5	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	0.1	< 0.1	< 0.1
credit-a	4.1	6.6	< 0.1	< 0.1	0.1	0.2	< 0.1	0.2	< 0.1	< 0.1
credit-g	18.8	40.8	< 0.1	0.1	0.2	0.9	0.1	0.9	< 0.1	0.1
diabetes	0.5	0.5	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
elevators	672.1	309.0	0.2	0.9	1.0	2.4	1.9	4.1	0.3	0.9
flare	0.3	0.4	< 0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	< 0.1	< 0.1
forestfires	2.0	2.9	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
glass	0.2	0.3	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
heart-c	1.3	1.8	< 0.1	< 0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	< 0.1
house	1149.7	464.3	0.2	1.5	0.8	6.1	1.5	7.3	0.6	1.3
housing	2.5	3.5	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
letter	436.6	275.2	0.1	1.3	0.2	3.9	0.7	10.5	0.5	1.5
mv	76.0	27.5	1.0	1.7	1.9	1.4	0.9	2.8	0.9	1.6
pole	430.7	447.8	26.8	1.6	263.0	38.6	186.8	81.0	0.3	2.5
sonar	2971.2	9084.8	< 0.1	< 0.1	< 0.1	0.1	< 0.1	0.4	< 0.1	< 0.1
spambase	4626.0	>4 h	692.9	609.5	7657.5	4269.1	3319.2	1900.3	0.2	151.2
ticdata	>4 h	>4 h	820.4	56.9	12686.5	>4 h	5588.2	>4 h	0.1	1.0
yeast	0.5	0.7	< 0.1	< 0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	< 0.1

times of the algorithms without optimistic estimate pruning shown in Table 5 reflect the actual algorithm runtimes for these measures.

6.3 Runtimes of the Full Algorithms

Another series of experiments compared the runtimes of the full algorithms. Exemplary results of these evaluations are shown in Table 6 and Table 7. Experiments depicted in Table 6 utilized different interestingness measures and a fixed search depth of $d = 5$ in a top-1 search . By contrast, experiments shown in Table 7 employed the fixed interestingness measure $q^{0.5}_{mean}$, but a variable search depth.

The results in Table 6 indicate that for a search depth of $d = 5$, the application of optimistic estimate bounds leads to a substantial reduction of runtimes in comparison to the variations without optimistic estimate pruning in almost all cases, cf. Table 5. The largest improvements can be observed for the interestingness measures q^1_{mean} and q^0_{mean}. This corresponds the respective reduction in necessary candidate evaluations, see Table 3. Although the applied pruning bounds are in theory tighter for the *NumBSD* algorithm,

Table 7 Comparison of the full algorithms: the table shows the runtimes in seconds of *NumBSD* (BSD), and *SD-Map** (SDM) with *all pruning options enabled* for different maximum search depths d. As interestingness measure, the mean test $q^{0.5}_{mean}$ was used.

Dataset	$d=2$		$d=3$		$d=4$		$d=5$		$d=6$	
	BSD	SDM	BSD	SDM	BSD	SDM	BSD	SDM	BSD	SDM
adults	2.5	3.7	5.0	4.3	8.4	5.1	11.1	6.1	12.5	7.0
ailerons	3.1	4.6	7.9	5.8	20.9	9.1	45.3	17.7	78.4	41.3
autos	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	<0.1	< 0.1	0.1	< 0.1	0.1
breast-w	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
census-kdd	92.2	96.6	353.6	199.1	1490.7	462.8	5184.3	1110.2	> 4 h	3469.5
communities	0.2	4.0	0.2	4.0	0.2	4.1	0.1	4.3	0.2	4.4
concrete_data	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
credit-a	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.2
credit-g	0.1	0.3	0.1	0.5	0.2	0.7	0.2	0.9	0.2	1.0
diabetes	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
elevators	0.6	2.1	0.8	2.2	0.9	2.2	1.0	2.4	1.1	2.3
flare	< 0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	0.1	< 0.1
forestfires	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
glass	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
heart-c	< 0.1	< 0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	0.1
house	0.7	5.3	0.7	5.5	0.7	5.6	0.8	6.1	0.7	5.6
housing	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	0.1
letter	0.2	3.7	0.2	3.8	0.2	3.8	0.2	3.9	0.2	3.8
mv	1.7	1.4	1.7	1.2	1.7	1.2	1.9	1.4	1.7	1.3
pole	3.5	2.7	18.6	5.9	79.8	14.7	263.0	38.6	699.2	90.2
sonar	< 0.1	0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	0.1
spambase	6.0	21.2	87.6	131.2	865.4	781.6	7657.5	4269.1	> 4 h	> 4 h
ticdata	10.0	88.5	125.9	539.1	1429.2	2939.1	12686.5	> 4 h	> 4 h	> 4 h
yeast	< 0.1	0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	0.1	< 0.1	0.1

the speedups are often larger for the *SD-Map** algorithm. This has two reasons: first, pruning in *SD-Map** is exploited twice, as header pruning and as selector pruning when conditional trees are built. Since the size of the conditional trees is reduced by pruning, not only fewer candidate evaluations are required with optimistic estimates, but each candidate evaluation also takes less time to compute. Second, the computation of the ordering-based bounds in *NumBSD* is more costly in itself. In one single experiment with unfavorable pruning properties, the computational costs for determining the bounds in this algorithm outweighed the use of pruning, that is, for the *spambase* dataset and the interestingness measure $q^{0.5}_{mean}$. For the interestingness measure q^{0}_{mean}, the runtimes of *SD-Map** are surprisingly high in the datasets *spambase* and *aileron*. This could be explained by the fact that for this interestingness measure and these datasets very specific subgroups have to be found early to exploit the optimal bounds. Additionally many ties occur in the sorting based on the optimistic estimates, which are differently solved in the different algorithm implementations. Therefore, *SD-Map** explores more candidates than necessary in the best case. However, there is still a substantial speedup in comparison to unpruned algorithm variants.

The runtimes of both proposed algorithms differ significantly in several cases. Unfortunately, a recommendation for choosing between the two novel

algorithms for a certain task remains difficult. As a tendency, *SD-Map** is preferred for the interestingness measures that select subgroups with higher coverage (q^{1}_{mean} and $q^{0.5}_{mean}$) if the runtime is not very short (< 5 seconds) anyway. On the other hand, *NumBSD* is in general faster for $q^{0.1}_{mean}$ and q^{0}_{mean}. However, there are several exceptions for this rule of thumb.

Table 7 displays the algorithm runtimes for different search depth using the interestingness measure $q^{0.5}_{mean}$. A comparison to the unpruned algorithms, see Table 5, shows substantial runtime improvements in most cases. The improvements were in particular strong for larger search depths, i.e., $d = 5$ and $d = 6$, where most runtimes decreased by more than an order of magnitude. For several datasets, the runtime did not (or only marginally) increase with higher search depths, e.g., for the datasets *autos*, *communities*, or *elevators*. This is a sharp contrast to the variants which do not employ optimistic estimate pruning and can be explained by the fact that already at search level two or three all further candidates can be pruned. Also for medium search depths $d = 3$ and $d = 4$, substantial runtime improvements can be observed in most cases with runtimes > 5 seconds, but the performance gains are not as large as for the high depth searches. For the minimum search depth $d = 2$, the gains for *SD-Map** were only moderate, while *NumBSD* took even more time than its variation without pruning. At this low search depth, the effects of the pruning seems to have less influence than the additional computational costs for computing the bounds. However, for this search depth the runtimes of *NumBSD* were very low in most cases anyway. For a few datasets the costs for computing the optimistic estimates exceeded the gains from utilizing the pruning bounds even for higher search depth, e.g., in the datasets *ticdata* and *spambase*. This was never the case for the *SD-Map** algorithm.

Comparing both novel algorithms with each other, *SD-Map** excels for higher search depths ($d \geq 4$), where it outperforms *NumBSD* for most experiments with relevant runtimes, that is, if tasks take more than five seconds to complete. By contrast, for the lower search depths $d = 3$ and especially $d = 2$, *NumBSD* performs better. In these cases, the overhead necessary for the FP-trees in *SD-Map** seems to be too high to be worth it. These results are in general in line with the previous recommendations for the unpruned algorithm versions. However, since *SD-Map** does profit more from the optimistic estimate bounds than *NumBSD*, it also performs better at the medium search depths $d = 3$ and $d = 4$ in several cases. The runtimes and thus the preferences of the algorithms do not correlate as strongly with the dataset size as in the unpruned variants, but also depend strongly on the pruning opportunities in the respective datasets. Unfortunately, the respective properties are difficult to determine beforehand.

In additional experiments, similar runtime improvements as for the mean-based interestingness measure could also be observed for other interestingness measures, such as median-based measures. As observed for the mean-based measures, the actual algorithm runtimes are highly correlated with the number of required candidate evaluations for the respective measure, cf. Table 4.

Table 8 Comparison of candidate evaluations for different sizes of the result set k if ordering-based optimistic estimate pruning is applied. As interestingness measure, several mean-based interestingness measures q^a_{mean} were used. The maximum search depth (maximum number of selectors in a description) was limited to $d = 5$ (due to long runtimes $d = 4$ for the datasets marked with "*").

Dataset		$q^{1.0}_{mean}$			$q^{0.5}_{mean}$			$q^{0.1}_{mean}$			$q^{0.0}_{mean}$		
	any k	$k=1$	$k=10$	$k=100$	$k=1$	$k=10$	$k=100$	$k=1$	$k=10$	$k=100$	$k=1$	$k=10$	$k=100$
adults	8,503,218	253	990	6,453	1,872	4,793	21,854	32,873	50,811	72,001	1,252	3,260	14,853
ailerons	984,289,405	4,298	4,890	13,055	4,384	5,440	18,935	148,830	173,320	223,993	1,470	2,177	8,426
autos	12,316,190	17	159	3,685	66	291	4,283	1,884	2,802	6,032	134	324	1,047
breast-w	219,993	12	86	3,710	19	178	3,751	963	1,702	3,819	132	206	889
census-kdd*	73,374,193	1,815	3,413	10,291	19,012	22,474	42,815	145,315	169,734	196,457	1,554	6,270	34,244
communities*	$> 2 \cdot 10^9$	22	290	10,849	92	1,266	32,852	79,187	126,796	314,546	2,248	3,167	3,664
concrete_data	209,041	19	336	4,744	131	866	4,294	707	971	4,034	457	564	3,921
credit-a	2,231,118	277	1,497	11,169	940	6,241	18,982	867	3,049	11,551	777	2,577	11,119
credit-g	8,389,271	364	1,767	16,375	906	8,703	24,604	383	490	1,793	319	414	1,743
diabetes	350,466	17	351	8,891	102	2,971	8,001	316	430	4,153	342	434	3,935
elevators	121,983,859	62	380	8,719	388	2,408	8,478	2,389	3,090	5,927	1,133	1,863	5,729
flare	101,946	29	227	1,692	446	1,210	3,062	1,497	2,487	5,402	5	84	215
forestfires	1,209,242	30	303	3,146	172	330	2,128	164	411	2,051	161	415	2,051
glass	141,714	10	143	4,750	17	360	4,678	87	168	1,157	147	233	1,138
heart-c	823,995	224	1,077	5,772	975	4,248	9,203	516	648	1,654	357	537	1,490
house	173,768,450	21	225	11,534	482	4,298	27,140	13,424	18,544	35,004	769	832	2,387
housing	1,554,972	43	116	81,065	387	688	81,706	273	430	80,457	127	427	80,171
letter	69,157,431	8	381	14,798	9	1,098	20,229	3,409	5,797	13,854	988	1,678	4,852
mv	5,542,943	34	1,299	9,450	35	2,073	9,655	34	1,524	10,508	1,499	6,122	33,687
pole	47,553,142	48,077	52,890	3,378,458	72,851	108,189	3,416,279	146,175	196,301	3,664,015	270	374	3,287,774
sonar	1,737,064,885	12	219	8,586	12	2,094	12,718	1,253	2,297	4,215	1,208	2,706	3,632
spambase*	35,545,851	304,187	427,429	583,275	641,296	646,133	661,439	310,587	313,094	322,891	878	1,941	13,106
ticdata*	430,966,974	218,921	228,157	255,675	650,638	697,812	929,897	837,239	916,105	1,046,667	61	3,498	12,787
yeast	291,042	23	481	3,182	27	656	3,022	259	772	2,985	217	522	2,787

6.4 Influence of the Result Set Size

In a top-k approach, the size of the result set k influences the effects of optimistic estimate pruning. Since it is exploited that candidates receive for sure a lower score than the best k subgroups found so far, pruning can be applied less often and more candidate subgroups must be explored for larger values of k. In another series of experiments we studied the influence of the result set size k on the number of required subgroup evaluations. Table 8 shows the number of candidate evaluations that were performed in a depth-first-search with one level look-ahead for different mean-based interestingness measures and different sizes k of the result set. The search depth was limited to $d = 5$ ($d = 4$ for some datasets with high runtimes). For pruning, ordering-based bounds were applied.

The results show that the number of evaluated candidates increases with the size of the result set. Nonetheless, the number of evaluated candidates is still smaller by orders of magnitudes than in a search without optimistic estimates. Fortunately, the (relative) increase is much more moderate for datasets that require large numbers of subgroup evaluations even with optimistic estimate pruning, see the datasets *spambase* and *ticdata*. This can be explained by the fact that the large amount of evaluations is required, because the dataset contains many subgroups with similar scores according to the applied interestingness measure. In this case, the number of required evaluations is also less influenced by the size of the result set k.

Additionally, we also tested the runtime of the full algorithms for higher settings of k. Table 9 shows the runtimes for $k = 100$. In comparison with the previous results, see Table 6, the runtimes are in many cases only marginally increased. However, for some datasets and interestingness measures the algorithms take significantly more time. In particular, for the average interestingness measure $q^{0.0}_{mean}$ the runtimes for *SD-Map** are increased in comparison to a top-1 search, see for example the datasets *census-kdd* or *ticdata*. Nonetheless, the algorithms are still substantially faster than their counterparts that do not employ optimistic estimate pruning. Overall, the presented optimistic estimate bounds are also clearly useful for larger result set sizes.

6.5 Effects of the Fast Pruning Bounds

Section 5.2.3 introduced a new category of optimistic estimates that can already be applied if only a part of the current subgroup is analyzed. These are incorporated in the *NumBSD* algorithm and were also included in the previous experiments. To measure the effects of the novel bounds, the runtimes of the full *NumBSD* algorithm was compared with a variation that did not employ these bounds. The search employed a maximum search depth of $d = 5$ and differently parametrized mean-based interestingness measures. The results are shown in Table 10. Datasets, for which the tasks could be solved very fast (< 0.2 seconds) by both variants, are omitted.

Table 9 Comparison of the full algorithms with a larger result set size: the table shows the runtimes in seconds for *NumBSD* (BSD) and for *SD-Map** (SDM) with *all pruning options enabled* for different interestingness measures. The search was limited to a maximum search depth of $d = 5$ and used a result set size of $k = 100$.

Dataset	q^1_{mean}		$q^{0.5}_{mean}$		$q^{0.1}_{mean}$		q^0_{mean}	
	BSD	SDM	BSD	SDM	BSD	SDM	BSD	SDM
adults	3.1	4.3	24.6	8.4	32.8	15.8	1.7	9.0
ailerons	6.8	2.4	57.7	19.9	219.7	3,379.0	0.7	459.1
autos	< 0.1	< 0.1	0.1	0.1	0.1	0.3	< 0.1	0.1
breast-w	< 0.1	< 0.1	0.1	0.1	0.1	0.1	< 0.1	0.1
census-kdd	122.8	71.3	5,505.0	1,150.7	> 4 h	5,914.2	36.1	1,876.5
communities	0.3	0.8	1.6	8.8	7.2	184.3	0.3	1.0
concrete_data	< 0.1	< 0.1	0.1	0.1	< 0.1	0.1	< 0.1	0.1
credit-a	0.1	0.1	0.3	0.3	0.1	0.3	0.1	0.1
credit-g	0.2	0.1	1.3	1.9	0.1	1.3	< 0.1	0.8
diabetes	< 0.1	< 0.1	0.1	0.1	< 0.1	0.1	< 0.1	< 0.1
elevators	0.7	1.4	3.6	2.9	3.9	4.5	0.4	1.5
flare	< 0.1	0.1	0.2	0.1	0.2	0.1	< 0.1	0.1
forestfires	< 0.1	< 0.1	< 0.1	0.1	< 0.1	0.1	< 0.1	< 0.1
glass	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
heart-c	< 0.1	< 0.1	0.2	0.2	< 0.1	0.1	< 0.1	0.1
house	0.9	2.7	3.4	9.9	2.7	8.7	0.6	2.3
housing	< 0.1	< 0.1	0.1	0.1	< 0.1	0.1	< 0.1	0.1
letter	0.8	2.6	2.8	7.9	1.1	11.8	0.5	3.4
mv	2.0	2.9	8.0	3.2	2.9	4.0	1.1	3.3
pole	29.0	2.4	273.9	40.5	280.3	95.8	0.3	7.6
sonar	0.1	0.1	0.2	0.3	0.1	0.5	0.1	< 0.1
spambase	701.2	618.6	7,708.3	4,306.6	3,338.9	1,906.4	0.2	812.7
ticdata	843.7	58.9	12,880.8	> 4 h	6,354.5	> 4 h	1.1	2,030.8
yeast	< 0.1	0.1	0.2	0.1	0.1	0.1	< 0.1	0.1

The results indicate that the influence of the additional bounds that can be computed early in the evaluation process is somewhat limited. The runtimes are most improved for the interestingness measure q^1_{mean}: For this measure, the improvements for most datasets are between 10% and 40%. Pruning bounds for this measure seem to be more easily exploitable since this measure requires subgroups that cover many instances.

For other interestingness measure the benefits are less significant: they do not exceed 10% in many cases. However, only in a single setting (for the dataset *mv*), the computational efforts of determining the additional bounds were higher than the saved efforts. Potentially, this kind of pruning requires additional optimization in the implementations to show its full benefits, e.g., by checking the additional bounds only at certain points in the evaluation.

Overall, it has to be reported that for now the novel advanced ("fast") type of bounds does not have the decisive effect yet. Instead, it is more of a minor addition in order to optimize the algorithm. However, in the future, this kind of pruning could be exploited with possibly stronger effects in distributed subgroup mining: if nodes are assigned to computational units according to their target values, and pruning bounds can already be applied at one unit,

then the other units are not required to be involved, thus potentially reducing the overall communication costs significantly.

Table 10 Evaluation of the full *NumBSD* (BSD) algorithm with a variation that does *not* employ the fast pruning bounds that can already be exploited by evaluating a part of the subgroup (NoFP). The comparison was performed with a maximum search depth of $d = 5$ and different mean-based interestingness measures. Datasets, for which the tasks could be solved very fast (< 0.2 seconds) by both variants, are omitted.

Dataset	q^1_{mean}		$q^{0.5}_{mean}$		$q^{0.1}_{mean}$		q^0_{mean}	
	NoFP	BSD	NoFP	BSD	NoFP	BSD	NoFP	BSD
adults	1.7	0.8	14.4	11.1	27.4	19.7	1.4	1.3
ailerons	6.5	6.0	47.1	45.3	110.9	95.5	0.6	0.6
census-kdd	89.6	53.3	5872.8	5184.3	>4 h	>4 h	35.6	35.0
communities	0.2	0.1	0.4	0.1	3.9	2.0	0.3	0.3
credit-g	0.1	< 0.1	0.4	0.2	0.1	0.1	< 0.1	< 0.1
elevators	0.3	0.2	1.3	1.0	3.2	1.9	0.3	0.3
house	0.5	0.2	1.2	0.8	3.2	1.5	0.6	0.6
letter	0.4	0.1	0.7	0.2	1.0	0.7	0.5	0.5
mv	0.7	1.0	1.3	1.9	1.0	0.9	0.8	0.9
pole	34.0	26.8	282.6	263.0	204.8	186.8	0.3	0.3
spambase	853.1	692.9	7989.2	7657.5	3555.2	3319.2	0.2	0.2
ticdata	1019.4	820.4	>4 h	12686.5	6440.6	5588.2	0.4	0.1

6.6 Evaluation Summary

The experiments clearly showed the effectiveness of the proposed improvements. The presented optimistic estimates were able to substantially reduce the number of required candidate evaluations for almost all interestingness measures. As expected, ordering-based bounds had even stronger effects, but bounds in closed forms were good approximations most of the time. Regarding data structures, both novel data structures outperformed a simple approach by far. While for searches with high search depths and large datasets the FP-tree structure enabled faster completion of the tasks, a bitset-based structure is better suited for the other tasks. A comparison of the full algorithms showed that improvements on data structures and optimistic estimate bounds can be combined well. The incorporation of the bounds further reduced the runtimes by an order of magnitude. The *SD-Map** algorithm did profit more from the additional pruning bounds since also the computational costs for single candidate evaluations are reduced. Although increasing the size of the result set reduces pruning possibilities, still the vast majority of the search space can be pruned in most cases. Unfortunately, a clear recommendation between the two novel algorithms remains difficult. As a tendency, *SD-Map** is to be preferred for more demanding tasks with higher search depths, while *NumBSD* performs better for low search depths.

7 Conclusions

In this paper, we investigated efficient exhaustive subgroup discovery with numerical target concepts. In order to provide a broad overview, we first surveyed interestingness measures for this setting from literature. These included mean-based, variance-based, median-based, and rank-based interestingness measures as well as a measure based on the Kolmogorov-Smirnov statistical test.

After that, we presented novel techniques to enable efficient exhaustive mining: we presented the adaptation of efficient data structures for the numerical target setting, that is, FP-trees and bitset-based data structures. Additionally, we investigated optimistic estimate bounds for pruning the search space and derived novel bounds for the discussed interestingness measures. In this context, we introduced *ordering-based bounds* as a flexible formalism that allows to derive optimistic estimates for interestingness measures with no previously known bounds. Additionally, we presented fast bounds that require only limited information about a subgroup and bounds in closed form that can also be determined in FP-trees. The proposed techniques, i.e., data structures and optimistic estimates, were incorporated in two novel algorithms, *SD-Map** and *NumBSD*. Using these, we provided an extensive experimental evaluation with 24 publicly available datasets. As a result, both novel algorithms outperformed simple approaches by orders of magnitudes. Possible advantages of one algorithm over the other were discussed.

For future work, a comparison of exhaustive and heuristic search algorithms in terms of runtime and result quality is planned. Additionally, the integration of methods for numeric attributes in the search space, see (Grosskreutz and Rüping 2009; Mampaey et al. 2012), and numeric target concepts will be of high relevance for practical applications. Furthermore, case studies that show the advantages and disadvantages of the discussed interestingness measures seem like an interesting direction for future research. Finally, we aim to investigate the option of generalized optimistic estimates for an extended view on subgroup discovery techniques, e. g., considering exceptional model mining (Leman et al. 2008; Duivesteijn et al. 2010; Lemmerich et al. 2012; Atzmueller et al. 2015), and generalization-aware methods (Lemmerich and Puppe 2011; Lemmerich et al. 2013).

Acknowledgements

This work has been partially supported by the VENUS research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University.

References

Alcala-Fernandez, J., Fernandez, A., Luengo, J., Derrac, J., Garcia, S., Sanchez, L., and Herrera, F. (2011). KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2–3):255–287.

Atzmueller, M. (2015). Subgroup Discovery – Advanced Review. *WIREs: Data Mining and Knowledge Discovery*, 5(1):35–49.

Atzmueller, M. and Lemmerich, F. (2009). Fast Subgroup Discovery for Continuous Target Concepts. In: *Proceedings of the 18th International Symposium on Foundations of Intelligent Systems (ISMIS)*, pp 35–44.

Atzmueller, M. and Lemmerich, F. (2012). VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pp 842–845.

Atzmueller, M. and Lemmerich, F. (2013). Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. *International Journal of Web Science*, 2(1–2):80–112.

Atzmueller, M., Lemmerich, F., Krause, B., and Hotho, A. (2009). Who are the Spammers? Understandable Local Patterns for Concept Description. In: *Proceedings of the 7th Conference on Computer Methods and Systems*.

Atzmueller, M., Mueller, J., and Becker, M. (2015). Exploratory Subgroup Analytics on Ubiquitous Data. In: Atzmueller, A., Chin, A., Scholz, C., and Trattner, C.(ed), *Mining, Modeling and Recommending 'Things' in Social Media*, pp 1–20.

Atzmueller, M. and Puppe, F. (2006). SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In: *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pp 6–17.

Atzmueller, M. and Puppe, F. (2009). A Knowledge-Intensive Approach for Semi-Automatic Causal Subgroup Discovery. In: Berendt, B. et al.(ed), *Knowledge Discovery Enhanced with Semantic and Social Information*, volume 220, pp 19–36.

Aumann, Y. and Lindell, Y. (1999). A Statistical Theory for Quantitative Association Rules. In: *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp 261–270.

Aumann, Y. and Lindell, Y. (2003). A Statistical Theory for Quantitative Association Rules. *Journal of Intelligent Information Systems*, 20(3):255–283.

Batal, I. and Hauskrecht, M. (2010). A Concise Representation of Association Rules using Minimal Predictive Rules. In: *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pp 87–102.

Bay, S. D. and Pazzani, M. J. (2001). Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 5(3):213–246.

Bayardo, R. J. (1998). Efficiently Mining Long Patterns from Databases. In: *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp 85–93.

Bayardo, R. J., Agrawal, R., and Gunopulos, D. (1999). Constraint-Based Rule Mining in Large, Dense Databases. *Data Mining and Knowledge Discovery*, 4(2–3):217–240.

Box, G. E. P. (1953). Non-Normality and Tests on Variances. *Biometrika*, 40:318–335.

Breiman, L., Friedman, J. H., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman & Hall.

Brin, S., Rastogi, R., and Shim, K. (2003). Mining Optimized Gain Rules for Numeric Attributes. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):324–338.

Cheng, H., Yan, X., Han, J., and Yu, P. S. (2008). Direct Discriminative Pattern Mining for Effective Classification. In: *Proceedings of the 24th International Conference on Data Engineering (ICDE)*, pp 169–178.

Dong, G. and Li, J. (1999). Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp 43–52.

Duivesteijn, W., Knobbe, A. J., Feelders, A., and van Leeuwen, M. (2010). Subgroup Discovery Meets Bayesian Networks – An Exceptional Model Mining approach. In: *Proceedings of the 10th International Conference on Data Mining (ICDM)*, pp 158–167.

El-Qawasmeh, E. (2003). Beating the Popcount. *International Journal of Information Technology*, 9(1):1–18.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fayyad, U. M. and Irani, K. B. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI)*, pp 1022–1027.

Freidlin, B. and Gastwirth, J. L. (2000). Should the Median Test be Retired from General Use? *The American Statistician*, 54(3):161–164.

Fukuda, T., Morimoto, Y., Morishita, S., and Tokuyama, T. (1996). Mining Optimized Association Rules for Numeric Attributes. In: *Proceedings of the 15th ACM Symposium on Principles of Database Systems (PODS)*, pp 182–191.

García, S., Luengo, J., Saez, J. A., Lopez, V., and Herrera, F. (2013). A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750.

Geng, L. and Hamilton, H. J. (2006). Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys*, 38(3):Article no. 9.

Grosskreutz, H. (2008). Cascaded Subgroups Discovery with an Application to Regression. In: *From Local Patterns to Global Models, Workshop at the ECML/PKDD*, pp 275–286.

Grosskreutz, H. and Rüping, S. (2009). On Subgroup Discovery in Numerical Domains. *Data Mining and Knowledge Discovery*, 19(2):210–226.

Grosskreutz, H., Rüping, S., and Wrobel, S. (2008). Tight Optimistic Estimates for Fast Subgroup Discovery. In: *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pp 440–456.

Han, J., Pei, J., and Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. *ACM SIGMOD Record*, 29(2):1–12.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery*, 8(1):53–87.

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

Jorge, A. M., Azevedo, P. J., and Pereira, F. (2006). Distribution Rules with Numeric Attributes of Interest. In: *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pp 247–258.

Kavšek, B. and Lavrač, N. (2006). Apriori-SD: Adapting Association Rule Learning To Subgroup Discovery. *Applied Artificial Intelligence*, 20:543–583.

Klösgen, W. (1994). Exploration of Simulation Experiments by Discovery. Technical Report WS-04-03.

Klösgen, W. (1995). Efficient Discovery of Interesting Statements in Databases. *Journal of Intelligent Information Systems*, 4(1):53–69.

Klösgen, W. (1996). Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R.(ed), *Advances in Knowledge Discovery and Data Mining*, pp 249–271.

Klösgen, W. (2002). Data Mining Tasks and Methods: Subgroup Discovery: Deviation Analysis. In: Klösgen, W. and Zytkow, J. M.(ed), *Handbook of Data Mining and Knowledge Discovery*, pp 354–361.

Klösgen, W. and May, M. (2002). Census Data Mining - An Application. In: *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*.

Kotsiantis, S. and Kanellopoulos, D. (2006). Discretization Techniques: A Recent Survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58.

Kralj Novak, P., Lavrač, N., and Webb, G. I. (2009). Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *Journal of Machine Learning Research*, 10:377–403.

Lavrač, N., Kavšek, B., Flach, P. A., and Todorovski, L. (2004). Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188.

Leman, D., Feelders, A., and Knobbe, A. J. (2008). Exceptional Model Mining. In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pp 1–16.

Lemmerich, F. (2014). *Novel Techniques for Efficient and Effective Subgroup Discovery.* PhD thesis, Universität Würzburg.

Lemmerich, F. and Atzmueller, M. (2012). Describing Locations using Tags and Images: Explorative Pattern Mining in Social Media. In: *Revised selected papers from the Workshops on Modeling and Mining Ubiquitous Social Media*, pp 77–96.

Lemmerich, F., Becker, M., and Atzmueller, M. (2012). Generic Pattern Trees for Exhaustive Exceptional Model Mining. In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pp 277–292.

Lemmerich, F., Becker, M., and Puppe, F. (2013). Difference-Based Estimates for Generalization-Aware Subgroup Discovery. In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pp 288–303.

Lemmerich, F. and Puppe, F. (2011). Local Models for Expectation-Driven Subgroup Discovery. In: *Proceedings of the 11th International Conference on Data Mining (ICDM)*, pp 360–369.

Lemmerich, F., Rohlfs, M., and Atzmueller, M. (2010). Fast Discovery of Relevant Subgroup Patterns. In: *Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp 428–433.

Lichman, M. (2013). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

Lucas, J. P., Jorge, A. M., Pereira, F., Pernas, A. M., and Machado, A. A. (2007). A Tool for Interactive Subgroup Discovery using Distribution Rules. In: *Proceedings of the Artificial Intelligence 13th Portuguese Conference on Progress in Artificial Intelligence (EPIA)*, pp 426–436.

Mampaey, M., Nijssen, S., Feelders, A., and Knobbe, A. J. (2012). Efficient Algorithms for Finding Richer Subgroup Descriptions in Numeric and Nominal Data. In: *Proceedings of the 12th International Conference on Data Mining (ICDM)*, pp 499–508.

Moreland, K. and Truemper, K. (2009). Discretization of Target Attributes for Subgroup Discovery. In: *Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, pp 44–52.

Morishita, S. (1998). On Classification and Regression. In: *Proceedings of the First International Conference on Discovery Science*, pp 40–57.

Morishita, S. and Sese, J. (2000). Traversing Itemset Lattices with Statistical Metric Pruning. In: *Proceedings of the 19th ACM Symposium on Principles of Database Systems (PODS)*, pp 226–236.

Pieters, B. F. I. (2010). Subgroup Discovery on Numeric and Ordinal Targets, with an Application to Biological Data Aggregation. Technical report, Universiteit Utrecht.

Pieters, B. F. I., Knobbe, A. J., and Džeroski, S. (2010). Subgroup Discovery in Ranked Data, with an Application to Gene Set Enrichment. In: *Preference Learning, Workshop at the ECML/PKDD*, volume 10, pp 1–18.

Rastogi, R. and Shim, K. (2002). Mining Optimized Association Rules with Categorical and Numeric Attributes. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):29–50.

Webb, G. I. (1995). OPUS: An Efficient Admissible Algorithm for Unordered Search. *Journal of Artificial Intelligence Research*, 3(1):431–465.

Webb, G. I. (2001). Discovering Associations with Numeric Variables. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp 383–388.

Wrobel, S. (1997). An Algorithm for Multi-relational Discovery of Subgroups. In: *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD)*, pp 78–87.

Zaki, M. J. (2000). Scalable Algorithms for Association Mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.

Zimmermann, A. and De Raedt, L. (2009). Cluster-Grouping: From Subgroup Discovery to Clustering. *Machine Learning*, 77(1):125–159.

Appendix

Lemma 1 *Using the notations of Theorem 9, the function* $f^a(x)(n+x)^a \cdot \left(\frac{\sigma+x\cdot\theta}{n+x} - \mu_\emptyset\right)$ *has no local maxima inside its domain of definition:*

$$f^a(x) \leq max(f(0), f(x_{max}))$$

Proof We distinguish three cases by the parameter a of the applied generic mean interestingness measure:

first, for $a = 1$, it holds that

$$\begin{aligned} f^1(x) &= (n+x)^1 \cdot \left(\frac{\sigma + x\cdot\theta}{n+x} - \mu_\emptyset\right) \\ &= \sigma + \theta x - \mu_\emptyset n - \mu_\emptyset x \\ &= (\theta - \mu_\emptyset)\cdot x + \sigma - \mu_\emptyset n \end{aligned}$$

As this is a linear function in x, the function $f^1(x)$ is strictly increasing for $\theta > \mu_\emptyset$ and strictly decreasing otherwise. Thus, the theorem holds for $a = 1$.

Second, we consider the case $(a \neq 1) \wedge (\sigma = \theta n)$, that is, the first n instances all had the same target value. In this case, the function $f^a(x)$ is given by $f^a(x) = (n+x)^a(\theta - \mu_\emptyset)$. This is strictly monotone since $n > 0, x > 0$. Thus, again $f^a(x)$ has no local maximum.

Third, the case $(a \neq 1) \wedge (\sigma \neq \theta n)$ is considered in detail: since σ was computed as a sum of n values that are at least as large as θ it can be assumed that $\theta \cdot n < \sigma$. In the following, the maxima of $f^a(x)$ is determined by deriving this function twice.

$$\begin{aligned} f^{a\,\prime}(x) = \frac{d}{dx} f^a(x) &= (n+x)^a \cdot \left(\frac{\sigma + x\cdot\theta}{n+x} - \mu_\emptyset\right) \\ &= (n+x)^a \left(\frac{d}{dx}\left(\frac{\theta x + \sigma}{n+x} - \mu_\emptyset\right)\right) + \left(\frac{\theta x+\sigma}{n+x} - \mu_\emptyset\right)\cdot\left(\frac{d}{dx}(n+x)^a\right) \\ &= (n+x)^a \left(\frac{d}{dx}\left(\frac{\theta x + \sigma}{n+x}\right)\right) + \left(\frac{\theta x+\sigma}{n+x} - \mu_\emptyset\right)\cdot a(n+x)^{a-1} \\ &= (n+x)^a \left(\frac{\theta}{n+x} - \frac{\theta x+\sigma}{(n+x)^2}\right) + \left(\frac{\theta x+\sigma}{n+x} - \mu_\emptyset\right)\cdot a(n+x)^{a-1} \\ &= (n+x)^{a-2}\left((\theta(n+x) - (\theta x+\sigma)) + a\,(\theta x + \sigma - \mu_\emptyset(n+x))\right) \\ &= (n+x)^{a-2}(\theta n - \sigma + a\theta x + a\sigma - a\mu_\emptyset n - a\mu_\emptyset x)) \\ &= (n+x)^{a-2}\left((x(a\theta - a\mu_\emptyset) + a\sigma - a\mu_\emptyset n + \theta n - \sigma\right) \end{aligned}$$

In line 2, the product rule is used. In line 3 the chain rule is applied, substituting (n+x). $\mu_\emptyset$ can be omitted, as it is constant with respect to x. In line 4 the quotient rule is used. Finally, in line 5 $(n+x)^{a-2}$ is factored out.

Since $x > 0, n > 0$ by definition, the first factor is obviously greater than zero for any valid x. For $a = 0$ or $\theta = \mu_\emptyset$, the second factor of this function is independent from x, so it has no root, thus $f(x)$ has no maxima except

the definition boundaries in this case. Otherwise the root of this function and therefore the only candidate for a maximum of $f^a(x)$ is given at the point

$$x^* = \frac{-a\sigma + an\mu_\emptyset - \theta n + \sigma}{a(\theta - \mu_\emptyset)}.$$

In the following, it is shown that x^* can not be a maximum value in our setting. For that purpose, the second derivative of $f(x)$ is computed at the point x^*:

$$\begin{aligned}
f^a{}''(x) &= \frac{d}{dx} f'(x) \\
&= (n+x)^{a-3}(a-2)(x(a\theta - a\mu_\emptyset) \\
&\quad + a\sigma - an + \theta n - \sigma) + (a\theta - a\mu_\emptyset)(n+x)^{a-2} \\
&= (n+x)^{a-3}((a-2)(x(a\theta - a\mu_\emptyset) \\
&\quad + a\sigma - an\mu_\emptyset + \theta n - \sigma) + (a\theta - a\mu_\emptyset)(n+x)) \\
&= (n+x)^{a-3}(a^2x\theta - a^2x\mu_\emptyset + a^2\sigma - a^2\mu_\emptyset n + a\theta n - a\sigma - 2xa\theta \\
&\quad + 2ax\mu_\emptyset - 2a\sigma + 2an\mu_\emptyset - 2\theta n + 2\sigma + a\theta n - an\mu_\emptyset + a\theta x - ax\mu_\emptyset) \\
&= (n+x)^{a-3}(a-1)(a\theta x + a\sigma - an\mu_\emptyset - ax\mu_\emptyset + 2\theta n - 2\sigma) \\
&= (n+x)^{a-3}(a-1)(x(a\theta - a\mu_\emptyset) + a\sigma - an\mu_\emptyset + 2\theta n - 2\sigma)
\end{aligned}$$

We now can determine the second derivative of f in x^*:

$$\begin{aligned}
f^a{}''(x^*) &= (n+x^*)^{a-3}(a-1)(x^*(a\theta - a\mu_\emptyset) + a\sigma - an\mu_\emptyset + 2\theta n - 2\sigma) \\
&= (n+x^*)^{a-3}(a-1) \\
&\quad \left(\frac{-a\sigma + an\mu_\emptyset - \theta n + \sigma}{a(\theta - \mu_\emptyset)}(a\theta - a\mu_\emptyset) + a\sigma - an\mu_\emptyset + 2\theta n - 2\sigma\right) \\
&= (n+x^*)^{a-3}(a-1)(-a\sigma + an\mu_\emptyset - \theta n + \sigma + a\sigma - an\mu_\emptyset + 2\theta n - 2\sigma) \\
&= (n+x^*)^{a-3}(a-1)(\theta n - \sigma) \\
&= (n+x^*)^{a-3}(a-1)(\theta n - \sigma)
\end{aligned}$$

Since $f^a(x)$ is defined only for positive x, the first factor is always positive. Since by premise $a < 1$ and $\theta n < \sigma$, the second derivative at point x^* is always positive. Thus, if x^* is an extreme value of $f(x)$, then it is a local minimum. Since it was shown above that $f(x)$ has no other candidates for extreme values besides x^*, this proves the lemma.

Lemma 2 *The generic mean-based measures q^a_{mean} are convex for $a = 1$ in the $(\sum T(c), i_P)$ space. They are not convex for arbitrary a.*

Proof For $a = 1$, the interestingness measure is given by $q^1_{mean}(P) = i_P \cdot (\mu_P - \mu_\emptyset) = i_P \cdot (\frac{\sum_{c \in P} T(c)}{i_P} - \mu_\emptyset) = \sum_{c \in P} T(c) - i_P \mu_\emptyset$. This function is linear in both $\sum T(C)$ and i_P. Since linear functions are known to be convex, q^1_{mean} is convex.

To show that generic mean-based measures are not convex in general, we show an example where the definition of convex for a function f(x), that is, $\forall x, y, \lambda \in (0,1) : f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$, is violated. In our case, x and y are each two-dimensional points in the $(\sum T(c), i_P)$ space. In that regard, we consider a dataset with $\mu_\emptyset = 0$ and the mean test interestingness measure $q^{0.5}_{mean}$. Then, the considered interestingness measure is given by $q^{0.5}_{mean} = i_P^{0.5} \cdot (\mu_P - \mu_\emptyset) = \frac{\sum_{c \in P} T(c)}{\sqrt{i_P}} := f(x)$. As two points in the $(\sum T(c), i_P)$ space for which the convexity condition is violated we choose $x = (-100, 2)$ and $y = (-100, 10)$. Additionally, we choose $\lambda = 0.5$. Then, the convexity inequality is violated:

$$
\begin{aligned}
f((1-\lambda)x + \lambda y) &\leq (1-\lambda)f(x) + \lambda f(y) \\
f((0.5x + 0.5y) &\leq 0.5 \cdot f(x) + 0.5 \cdot f(y) \\
f((-100, 6)) &\leq 0.5 \cdot f((-100, 2)) + 0.5 \cdot f(-100, 10) \\
\frac{-100}{\sqrt{6}} &\leq 0.5 \cdot \frac{-100}{\sqrt{2}} + 0.5 \cdot \frac{-100}{\sqrt{10}} \\
\approx -40.82 \leq &\quad \approx -51.17
\end{aligned}
$$

Since the definition of convexity is violated in at least one example, the mean test interestingness measure $q^{0.5}_{mean}$ is not convex.

The non-convexity of $q^{0.5}_{mean}$ is also evident by a surface plot of the function for $\mu_\emptyset = 0$, see Figure 3. □

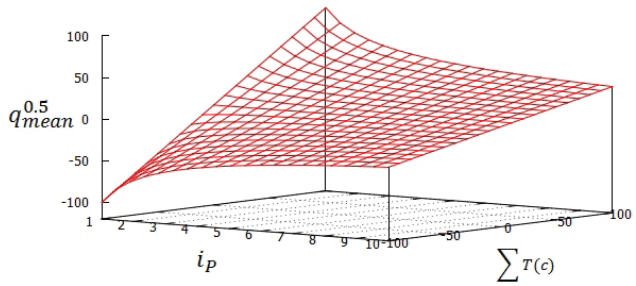

Fig. 3 A surface plot of the mean test interestingness measure $q^{0.5}_{mean}$ for $\mu_\emptyset = 0$ shows the non-convexity of this measure.