Data Min Knowl Disc
DOI 10.1007/s10618-017-0520-3

Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery

Mario Boley[1,2] · **Bryan R. Goldsmith**[2] · **Luca M. Ghiringhelli**[2] · **Jilles Vreeken**[1]

Received: 19 January 2017 / Accepted: 12 June 2017

Abstract Existing algorithms for subgroup discovery with numerical targets do not optimize the error or target variable dispersion of the groups they find. This often leads to unreliable or inconsistent statements about the data, rendering practical applications, especially in scientific domains, futile. Therefore, we here extend the optimistic estimator framework for optimal subgroup discovery to a new class of objective functions: we show how tight estimators can be computed efficiently for all functions that are determined by subgroup size (non-decreasing dependence), the subgroup median value, and a dispersion measure around the median (non-increasing dependence). In the important special case when dispersion is measured using the mean absolute deviation from the median, this novel approach yields a linear time algorithm. Empirical evaluation on a wide range of datasets shows that, when used within branch-and-bound search, this approach is highly efficient and indeed discovers subgroups with much smaller errors.

Keywords Subgroup discovery · Local pattern discovery · Branch-and-bound search

Responsible editors: Kurt Driessens, Dragi Kocev, Marko Robnik-Šikonja, Myra Spiliopoulou.

✉ Mario Boley
mboley@mpi-inf.mpg.de

Bryan R. Goldsmith
goldsmith@fhi-berlin.mpg.de

Luca M. Ghiringhelli
ghiringhelli@fhi-berlin.mpg.de

Jilles Vreeken
jilles@mpi-inf.mpg.de

[1] Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany

[2] Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Published online: 28 June 2017

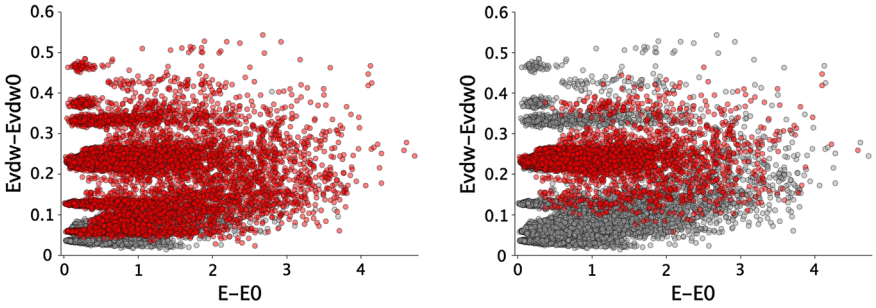
1 Introduction

Subgroup discovery is a well-established KDD technique (Klösgen 1996; Friedman and Fisher 1999; Bay and Pazzani 2001; see Atzmueller 2015 for a recent survey) with applications, e.g., in Medicine (Schmidt et al. 2010), Social Science (Grosskreutz et al. 2010), and Materials Science (Goldsmith et al. 2017). In contrast to global modeling, which is concerned with the complete characterization of some variable defined for a given population, subgroup discovery aims to detect intuitive descriptions of subpopulations in which, *locally*, the target variable has an interesting or useful distribution. In scientific domains, like the ones mentioned above, such local patterns are typically considered useful if they are not too specific (in terms of subpopulation size) and indicate insightful facts about the underlying physical process that governs the target variable. Such facts could for instance be: 'patients of specific demographics experience a low response to some treatment' or 'materials with specific atomic composition exhibit a high thermal conductivity'. For numeric (metric) variables, subgroups need to satisfy two criteria to truthfully represent such statements: the local distribution of the target variable must have a shifted central tendency (effect), and group members must be described well by that shift (consistency). The second requirement is captured by the group's *dispersion*, which determines the average error of associating group members with the central tendency value (see also Song et al. 2016).

Despite all three parameters—size, central tendency, and dispersion—being important, the only known approach for the efficient discovery of globally optimal subgroups, branch-and-bound search (Webb 1995; Wrobel 1997), is restricted to objective functions that only take into account size and central tendency. That is, if we denote by Q some subpopulation of our global population P then the objective functions f currently available to branch-and-bound can be written as

$$f(Q) = g(|Q|, c(Q)) \tag{1}$$

where c is some measure of central tendency (usually mean or median) and g is a function that is monotonically increasing in the subpopulation size $|Q|$. A problem with *all* such functions is that they inherently favor larger groups with scattered target values over smaller more focused groups with the same central tendency. That is, they favor the discovery of *inconsistent* statements over consistent ones—surprisingly often identifying groups with a local error that is almost as high or even higher than the global error (see Fig. 1 for an illustration of this problem that abounded from the authors' research in Materials Science). Although *dispersion-corrected* objective functions that counter-balance size by dispersion have been proposed (e.g., 't-score' by Klösgen 2002 or 'mmad' by Pieters et al. 2010), it remained unclear how to employ such functions outside of heuristic optimization frameworks such as greedy beam search (Lavrač et al. 2004) or selector sampling (Boley et al. 2012; Li and Zaki 2016). Despite often finding interesting groups, such frameworks do not guarantee the detection of optimal results, which can not only be problematic for missing important discoveries but also because they therefore can never guarantee the *absence* of high quality groups—which often is an insight equally important as the presence of a strong pattern. For instance, in our



(a) optimizing coverage times median shift; $\sigma_0(\cdot) \equiv (a(\cdot) \geq 6) \wedge (c_2(\cdot) > 0) \wedge (c_6(\cdot) <$ `v. high`); subgroup median 0.22, subgroup error 0.081

(b) optimizing dispersion-corrected variant; $\sigma_1(\cdot) \equiv (a(\cdot) \in [8, 12]) \wedge (c_2(\cdot) >$ `low`$) \wedge (c_6(\cdot) <$ `v. high`$) \wedge (r(\cdot) >$ `v. low`$)$; subgroup median 0.23, subgroup error 0.028

Fig. 1 To gain an understanding of the contribution of long-range van der Waals interactions (y-axis; above) to the total energy (x-axis; above) of gas-phase gold nanoclusters, subgroup discovery is used to analyze a dataset of such clusters simulated ab initio by density functional theory (Goldsmith et al. 2017); available features describe nanocluster geometry and contain, e.g., *number of atoms* a, *fraction of atoms with* i *bonds* c_i, and *radius of gyration* r. Here, similar to other scientific scenarios, a subgroup constitutes a useful piece of knowledge if it conveys a statement about a remarkable amount of van der Waals energy (captured by the group's central tendency) with high consistency (captured by the group's dispersion/error); optimal selector σ_0 with standard objective has high error and contains a large fraction of gold nanoclusters with a target value below the global median (0.13) (**a**); this is not the case for selector σ_1 discovered through dispersion-corrected objective (**b**), which therefore can be more consistently stated to describe gold nanoclusters with high van der Waals energy

example in Fig. 1, it would be remarkable to establish that long-range interactions are to a large degree independent of nanocluster geometry.

Therefore, in this paper (Sect. 3), we extend branch-and-bound search to objective functions of the form

$$f(Q) = g(|Q|, \mathtt{med}(Q), d(Q)) \tag{2}$$

where g is monotonically increasing in the subpopulation size, monotonically decreasing in any dispersion measure d around the median, and, besides that, depends only (but in arbitrary form) on the subpopulation median. This involves developing an efficient algorithm for computing the *tight optimistic estimator* given by the optimal value of the objective function among all possible subsets of target values:

$$\hat{f}(Q) = \max\{f(R) \colon R \subseteq Q\}\,, \tag{3}$$

which has been shown to be a crucial ingredient for the practical applicability of branch-and-bound (Grosskreutz et al. 2008; Lemmerich et al. 2016). So far, the most general approach to this problem (first codified in Lemmerich et al. (2016); generalized here in Sect. 3.1) is to maintain a sorted list of target values throughout the search process and then to compute Eq. (3) as the maximum of all subsets $R_i \subseteq Q$ that contain all target values of Q down to target value i—an algorithm that does not generalize to objective functions depending on dispersion. This paper presents an alternative idea (Sect. 3.2) where we do not fix the size of subset R_i as in the previous approach but instead fix its median to target value i. It turns out that this suffices

to efficiently compute the tight optimistic estimator for all objective functions of the form of Eq. (2). Moreover, we end up with a linear time algorithm (Sec. 3.3) in the important special case where the dependence on size and dispersion is determined by the *dispersion-corrected coverage* defined by

$$\texttt{dcc}(Q) = \frac{|Q|}{|P|} \max\left\{1 - \frac{\texttt{amd}(Q)}{\texttt{amd}(P)}, 0\right\}$$

where amd denotes the mean absolute deviation from the median. This is the same computational complexity as the objective function itself. Consequently, this new approach can discover subgroups according to a more refined selection criterion without increasing the worst-case computational cost. Additionally, as demonstrated by empirical results on a wide range of datasets (Sect. 4), it is also highly efficient and successfully reduces the error of result subgroups in practice.

2 Subgroup discovery

Before developing the novel approach to tight optimistic estimator computation, we recall in this section the necessary basics of optimal subgroup discovery with numeric target attributes. We focus on concepts that are essential from the optimization point of view (see, e.g., Duivesteijn and Knobbe 2011 and references therein for statistical considerations). As notional convention, we are using the symbol $[m]$ for a positive integer m to denote the set of integers $\{1, \ldots, m\}$. Also, for a real-valued expression x we write $(x)_+$ to denote $\max\{x, 0\}$. A summary of the most important notations used in this paper can be found in "Appendix C".

2.1 Description languages, objective functions, and closed selectors

Let P denote our given **global population** of entities, for each of which we know the value of a real **target variable** $y\colon P \to \mathbb{R}$ and additional descriptive information that is captured in some abstract **description language** $\mathcal{L}$ of subgroup selectors $\sigma\colon P \to \{\text{true}, \text{false}\}$. Each of these selectors describes a subpopulation $\mathbf{ext}(\sigma) \subseteq P$ defined by

$$\mathbf{ext}(\sigma) = \{p \in P\colon \sigma(p) = \text{true}\}$$

that is referred to as the **extension** of σ. Subgroup discovery is concerned with finding selectors $\sigma \in \mathcal{L}$ that have a useful (or interesting) distribution of target values in their extension $y_\sigma = \{y(p) : p \in \mathbf{ext}(\sigma)\}$. This notion of usefulness is given by an **objective function** $f : \mathcal{L} \to \mathbb{R}$. That is, the formal goal is to find selectors $\sigma \in \mathcal{L}$ with maximal $f(\sigma)$. Since we assume f to be a function of the multiset of y-values, let us define $f(\sigma) = f(\mathbf{ext}(\sigma)) = f(y_\sigma)$ to be used interchangeably for convenience. One example of a commonly used objective function is the **impact measure** ipa (see Webb 2001; here a scaled but order-equivalent version is given) defined by

$$\mathtt{ipa}(Q) = \mathtt{cov}(Q) \left(\frac{\mathtt{mean}(Q) - \mathtt{mean}(P)}{\mathtt{max}(P) - \mathtt{mean}(P)} \right)_+ \tag{4}$$

where $\mathtt{cov}(Q) = |Q|/|P|$ denotes the **coverage** or relative size of Q (here—and wherever else convenient—we identify a subpopulation $Q \subseteq P$ with the multiset of its target values).

The standard description language in the subgroup discovery literature[1] is the language $\mathcal{L}_{\text{cnj}}$ consisting of **logical conjunctions** of a number of base propositions (or predicates). That is, $\sigma \in \mathcal{L}_{\text{cnj}}$ are of the form

$$\sigma(\cdot) \equiv \pi_{i_1}(\cdot) \wedge \cdots \wedge \pi_{i_l}(\cdot)$$

where the π_{i_j} are taken from a pool of **base propositions** $\Pi = \{\pi_1, \ldots, \pi_k\}$. These propositions usually correspond to equality or inequality constraints with respect to one variable x out of a set of description variables $\{x_1, \ldots, x_n\}$ that are observed for all population members (e.g., $\pi(p) \equiv x(p) \geq v$). However, for the scope of this paper it is sufficient to simply regard them as abstract Boolean functions $\pi : P \to \{\text{true}, \text{false}\}$. In this paper, we focus in particular on the refined language of **closed conjunctions** $\mathcal{C}_{\text{cnj}} \subseteq \mathcal{L}_{\text{cnj}}$ (Pasquier et al. 1999), which is defined as $\mathcal{C}_{\text{cnj}} = \{\sigma \in \mathcal{L}_{\text{cnj}} : \mathbf{c}(\sigma) = \sigma\}$ by the fixpoints of the **closure operation** $\mathbf{c} \colon \mathcal{L}_{\text{cnj}} \to \mathcal{L}_{\text{cnj}}$ given by

$$\mathbf{c}(\sigma) = \bigwedge \{\pi \in \Pi : \mathbf{ext}(\pi) \supseteq \mathbf{ext}(\sigma)\} . \tag{5}$$

These are selectors to which no further proposition can be added without reducing their extension, and it can be shown that $\mathcal{C}_{\text{cnj}}$ contains at most one selector for each possible extension. While this can reduce the search space for finding optimal subgroups by several orders of magnitude, closed conjunctions are the longest (and most redundant) description for their extension and thus do not constitute intuitive descriptions by themselves. Hence, for reporting concrete selectors (as in Fig. 1), closed conjunctions have to be simplified to selectors of approximately minimum length that describe the same extension (Boley and Grosskreutz 2009).

2.2 Branch-and-bound and optimistic estimators

The standard algorithmic approach for finding optimal subgroups with respect to a given objective function is branch-and-bound search—a versatile algorithmic puzzle solving framework with several forms and flavors (see, e.g., Mehlhorn and Sanders 2008, Chap. 12.4). At its core, all of its variants assume the availability and efficient computability of two ingredients:

1. A **refinement operator** $\mathbf{r} : \mathcal{L} \to 2^{\mathcal{L}}$ that is monotone, i.e., for $\sigma, \varphi \in \mathcal{L}$ with $\varphi \in \mathbf{r}(\sigma)$ it holds that $\mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma)$, and that non-redundantly generates $\mathcal{L}$. That is, there is a root selector $\bot \in \mathcal{L}$ such that for every $\sigma \in \mathcal{L}$ there is a unique

[1] In this article we remain with this basic setting for the sake of simplicity. It is, however, important to note that several generalizations of this concept have been proposed (e.g., Parthasarathy et al. 1999; Huan et al. 2003), to which the contributions of this paper remain applicable.

```
Bst-BB(F, σ):            // F max. priority queue w.r.t. f̂, σ current f-maximizer
begin
    if F = ∅ or f̂(top(F))/f(σ) ≤ a then
        return σ
    else
        R = r(top(F))                    // refinement of f̂-maximizer in queue
        σ' = argmax(f(φ) : φ ∈ {σ} ∪ R)
        F' = (F \ {top(F)}) ∪ {φ ∈ R: f̂(φ)/f(σ') ≥ a}
        return Bst-BB(F', σ')
    end
end
σ* = Bst-BB({⊥}, ⊥)  // call with root element to find global solution
```

Algorithm 1: Best-first branch-and-bound that finds a-approximation to objective function f based on refinement operator $\mathbf{r}$ and optimistic estimator $\hat{f}$; depth-limit and multiple solutions (top-k) parameters omitted; `top` denotes the find max operation for priority queue.

sequence of selectors $\bot = \sigma_0, \sigma_1, \ldots, \sigma_l = \sigma$ with $\sigma_i \in \mathbf{r}(\sigma_{i-1})$. In other words, the refinement operator implicitly represents a directed tree (arborescence) on the description language $\mathcal{L}$ rooted in $\bot$.

2. An **optimistic estimator** (or bounding function) $\hat{f} : \mathcal{L} \to \mathbb{R}$ that bounds from above the attainable subgroup value of a selector among all more specific selectors, i.e., it holds that $\hat{f}(\sigma) \geq f(\varphi)$ for all $\varphi \in \mathcal{L}$ with $\mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma)$.

Based on these ingredients, a branch-and-bound algorithm simply enumerates all elements of $\mathcal{L}$ starting from $\bot$ using $\mathbf{r}$ (branch), but—based on $\hat{f}$—avoids expanding descriptions that cannot yield an improvement over the best subgroups found so far (bound). Depending on the order in which language elements are expanded, one distinguishes between depth-first, breadth-first, breadth-first iterative deepening, and best-first search. In the last variant, the optimistic estimator is not only used for pruning the search space, but also to select the next element to be expanded, which is particularly appealing for informed, i.e., *tight* optimistic estimators. An important feature of branch-and-bound is that it effortlessly allows to speed-up the search in a sound way by relaxing the result requirement from being f-optimal to just being an a**-approximation**. That is, the found solution σ satisfies for all $\sigma' \in \mathcal{L}$ that $f(\sigma)/f(\sigma') \geq a$ for some **approximation factor** $a \in (0, 1]$. The pseudo-code given in Algorithm 1 summarizes all of the above ideas. Note that, for the sake of clarity, we omitted here some other common parameters such as a depth-limit and multiple solutions (top-k), which are straightforward to incorporate (see Lemmerich et al. 2016).

An efficiently computable refinement operator has to be constructed specifically for the desired description language. For example for the language of conjunctions $\mathcal{L}_{\text{cnj}}$, one can define $\mathbf{r}_{\text{cnj}} : \mathcal{L}_{\text{cnj}} \to \mathcal{L}_{\text{cnj}}$ by

$$\mathbf{r}_{\text{cnj}}(\sigma) = \{\sigma \wedge \pi_i : \max\{j : \pi_j \in \sigma\} < i \leq k\}$$

where we identify a conjunction with the set of base propositions it contains. For the closed conjunctions $\mathbf{c}_{\text{cnj}}$, let us define the lexicographical prefix of a conjunction $\sigma \in \mathcal{L}_{\text{cnj}}$ and a base proposition index $i \in [k]$ as $\sigma|_i = \sigma \cap \{\pi_1, \ldots, \pi_i\}$. Moreover, let us denote with $\mathbf{i}(\sigma)$ the minimal index such that the i-prefix of σ is extension-preserving, i.e., $\mathbf{i}(\sigma) = \min\{i : \mathbf{ext}(\sigma|_i) = \mathbf{ext}(\sigma)\}$. With this we can construct a refinement operator (Uno et al. 2004) $\mathbf{r}_{\text{ccj}} : \mathcal{C}_{\text{cnj}} \to 2^{\mathcal{C}_{\text{cnj}}}$ as

$$\mathbf{r}_{\text{ccj}}(\sigma) = \{\varphi : \varphi = \mathbf{c}_{\text{cnj}}(\sigma \wedge \pi_j), \mathbf{i}(\sigma) < j \leq k, \pi_j \notin \sigma, \varphi|_j = \sigma|_j\}\,.$$

That is, a selector φ is among the refinements of σ if φ can be generated by an application of the closure operator given in Eq. (5) that is prefix-preserving.

How to obtain an optimistic estimator for an objective function of interest depends on the definition of that objective. For instance, the coverage function `cov` is a valid optimistic estimator for the impact function `ipa` as defined in Eq. (4), because the second factor of the impact function is upper bounded by 1. In fact there are many different optimistic estimators for a given objective function. Clearly, the smaller the value of the bounding function for a candidate subpopulation, the higher is the potential for pruning the corresponding branch from the enumeration tree. Ideally, one would like to use $\hat{f}(\sigma) = \max\{f(\varphi) : \mathbf{ext}(\varphi) \subseteq \mathbf{ext}(\sigma)\}$, which is the most strict function that still is a valid optimistic estimator. Computing this function, however, is as hard as the whole subgroup optimization problem. Thus, as a next best option, one can disregard subset selectability and consider the (selection-unaware) **tight optimistic estimator** (Grosskreutz et al. 2008) given by

$$\hat{f}(\sigma) = \max\{f(R) : R \subseteq \mathbf{ext}(\sigma)\}\,.$$

This leaves us with a new combinatorial optimization problem: given a subpopulation $Q \subseteq P$, find a sub-selection of Q that maximizes f. In the following section we will discuss strategies for solving this optimization problem efficiently for different classes of objective functions—including dispersion-corrected objectives.

3 Efficiently computable tight optimistic estimators

We are going to develop an efficient algorithm for the tight optimistic estimator in three steps: First, we review and reformulate a general algorithm for the classic case of non-dispersion-aware objective functions. Then we transfer the main idea of this algorithm to the case of dispersion-corrected objectives based on the median, and finally we consider a subclass of these functions where the approach can be computed in linear time. Throughout this section we will identify a given subpopulation $Q \subseteq P$ with the multiset of its target values $\{y_1, \ldots, y_m\}$ and assume that the target values are **indexed in ascending order**, i.e., $y_i \leq y_j$ for $i \leq j$. Also, for two multisets $Y = \{y_1, \ldots, y_m\}$ and $Z = \{z_1, \ldots, z_{m'}\}$ indexed in ascending order we say that Y is **element-wise less or equal** to Z and write $Y \preceq_e Z$ if $y_i \leq z_i$ for all $i \in [\min\{m, m'\}]$.

3.1 The standard case: monotone functions of a central tendency measure

The most general previous approach for computing the tight optimistic estimator for subgroup discovery with a metric target variable is described by Lemmerich et al. (2016), where it is referred to as *estimation by ordering*. Here, we review this approach and give a uniform and generalized version of that paper's results. For this, we define the general notion of a measure of central tendency as follows.

Definition 1 We call a mapping $c : \mathbb{N}^{\mathbb{R}} \to \mathbb{R}$ a (monotone) **measure of central tendency** if for all multisets $Y, Z \in \mathbb{N}^{\mathbb{R}}$ with $Y \preceq_e Z$ it holds that $c(Y) \leq c(Z)$.

One can check that this definition applies to the standard measures of central tendency, i.e., the arithmetic and geometric mean as well as the **median**[2] $\texttt{med}(Q) = y_{\lceil m/2 \rceil}$, and also to weighted variants of them (note, however, that it does not apply to the mode). With this we can define the class of objective functions for which the tight optimistic estimator can be computed efficiently by the standard approach as follows. We call $f : 2^P \to \mathbb{R}$ a **monotone level 1 objective function** if it can be written as

$$f(Q) = g(|Q|, c(Q))$$

where c is some measure of central tendency and g is a function that is non-decreasing in both of its arguments. One can check that the impact measure `ipa` falls under this category of functions as do many of its variants.

The central observation for computing the tight optimistic estimator for monotone level 1 functions is that the optimum value must be attained on a sub-multiset that contains a consecutive segment of elements of Q from the top element w.r.t. y down to some cut-off element. Formally, let us define the **top sequence** of sub-multisets of Q as $T_i = \{y_{m-i+1}, \ldots, y_m\}$ for $i \in [m]$ and note the following observation:

Proposition 1 *Let f be a monotone level 1 objective function. Then the tight optimistic estimator of f can be computed as the maximum value on the top sequence, i.e., $\hat{f}(Q) = \max\{f(T_i) : i \in [m]\}$.*

Proof Let $R \subseteq Q$ be of size k with $R = \{y_{i_1}, \ldots, y_{i_k}\}$. Since $y_{i_j} \leq y_{m-j+1}$, we have for the top sequence element T_k that $R \preceq_e T_k$ and, hence, $c(R) \leq c(T_k)$ implying

$$f(R) = g(k, c(R)) \leq g(k, c(T_l)) = f(T_k)\,.$$

It follows that for each sub-multiset of Q there is a top sequence element of at least equal objective value. □

From this insight it is easy to derive an $\mathcal{O}(m)$ algorithm for computing the tight optimistic estimator under the additional assumption that we can compute g and the "incremental central tendency problem" $(i, Q, (c(T_1), \ldots, c(T_{i-1})) \mapsto c(T_i)$ in constant time. Note that computing the incremental problem in constant time implies to

[2] In this paper, we are using the simple definition of the median as the 0.5-quantile (as opposed to defining it as $(y_{m/2} + y_{1+m/2})/2$ for even m), which simplifies many of the definitions below and additionally is well-defined in settings where averaging of target values is undesired.

only access a constant number of target values and of the previously computed central tendency values. This can for instance be done for $c = \texttt{mean}$ via the incremental formula $\texttt{mean}(T_i) = ((i-1)\,\texttt{mean}(T_{i-1}) + y_{m-i+1})/i$ or for $c = \texttt{med}$ through direct index access of either of the two values $y_{m-\lfloor (i-1)/2 \rfloor}$ or $y_{m-\lceil (i-1)/2 \rceil}$. Since, according to Proposition 1, we have to evaluate f only for the m candidates T_i to find $\hat{f}(Q)$ we can do so in time $\mathcal{O}(m)$ by solving the problem incrementally for $i = 1, \ldots, m$. The same overall approach can be readily generalized for objective functions that are monotonically decreasing in the central tendency or those that can be written as the maximum of one monotonically increasing and one monotonically decreasing level 1 function. However, it breaks down for objective functions that depend on more than just size and central tendency—which inherently is the case when we want to incorporate dispersion-control.

3.2 Dispersion-corrected objective functions based on the median

We will now extend the previous recipe for computing the tight optimistic estimator to objective functions that depend not only on subpopulation size and central tendency but also on the target value dispersion in the subgroup. Specifically, we focus on the median as measure of central tendency and consider functions that are both monotonically increasing in the described subpopulation size and monotonically decreasing in some dispersion measure around the median. To precisely describe this class of functions, we first have to formalize the notion of dispersion measure around the median. For our purpose the following definition suffices. Let us denote by $Y_\Delta^{\texttt{med}}$ the **multiset of absolute differences** to the median of a multiset $Y \in \mathbb{N}^{\mathbb{R}}$, i.e., $Y_\Delta^{\texttt{med}} = \{|y_1 - \texttt{med}(Y)|, \ldots, |y_m - \texttt{med}(Y)|\}$.

Definition 2 We call a mapping $d : \mathbb{N}^{\mathbb{R}} \to \mathbb{R}$ a **dispersion measure around the median** if $d(Y)$ is monotone with respect to the multiset of absolute differences to its median $Y_\Delta^{\texttt{med}}$, i.e., if $Y_\Delta^{\texttt{med}} \leq_e Z_\Delta^{\texttt{med}}$ then $d(Y) \leq d(Z)$.

One can check that this definition contains the measures median absolute deviation around the median $\texttt{mmd}(Y) = \texttt{med}(Y_\Delta^{\texttt{med}})$, the root mean of squared deviations around the median $\texttt{rsm}(Y) = \texttt{mean}(\{x^2 : x \in Y_\Delta^{\texttt{med}}\})^{1/2}$, as well as the **mean absolute deviation around the median** $\texttt{amd}(Y) = \texttt{mean}(Y_\Delta^{\texttt{med}})$.[3] Based on Def. 2 we can specify the class of objective functions that we aim to tackle as follows: we call a function $f : 2^P \to \mathbb{R}$ a **dispersion-corrected or level 2 objective function** (based on the median) if it can be written as

$$f(Q) = g(|Q|, \texttt{med}(Q), d(Q)) \tag{6}$$

[3] We work here with the given definition of dispersion measure because of its simplicity. Note, however, that all subsequent arguments can be extended in a straightforward way to a wider class of dispersion measures by considering the multisets of positive and negative deviations separately. This wider class also contains the interquartile range and certain asymmetric measures, which are not covered by Def. 2.

where d is some dispersion measure around the median and $g : \mathbb{R}^3 \to \mathbb{R}$ is a real function that is non-decreasing in its first argument and non-increasing in its third argument (without any monotonicity requirement for the second argument).

Our recipe for optimizing these functions is then to consider only subpopulations $R \subseteq Q$ that can be formed by selecting all individuals with a target value in some interval. Formally, for a fixed index $z \in \{1, \ldots, m\}$ define $m_z \leq m$ as the maximal cardinality of a sub-multiset of the target values that has median index z, i.e.,

$$m_z = \min\{2z, 2(m - z) + 1\}\,. \tag{7}$$

Now, for $k \in [m_z]$, let us define Q_z^k as the set with k *consecutive* elements around index z. That is

$$Q_z^k = \left\{ y_{z-\left\lfloor \frac{k-1}{2} \right\rfloor}, \cdots, y_z, \cdots, y_{z+\left\lceil \frac{k-1}{2} \right\rceil} \right\}\,. \tag{8}$$

With this we can define the elements of the **median sequence** Q_z as those subsets of the form of Eq. (8) that maximize f for some fixed index $z \in [m]$. That is, $Q_z = Q_z^{k_z^*}$ where $k_z^* \in [m_z]$ is minimal with

$$f(Q_z^{k_z^*}) = g(k_z^*, y_z, d(Q_z^{k_z^*})) = \max\{f(Q_z^k) \colon\ k \in [m_z]\}\,.$$

Thus, the number k_z^* is the smallest cardinality that maximizes the trade-off of size and dispersion encoded by g (given the fixed median $y_z = \texttt{med}(Q_z^k)$ for all k).

Figure 2 shows an exemplary median sequence based on 21 random target values. Note how the set sizes k_z^* vary non-monotonically for increasing median indices z

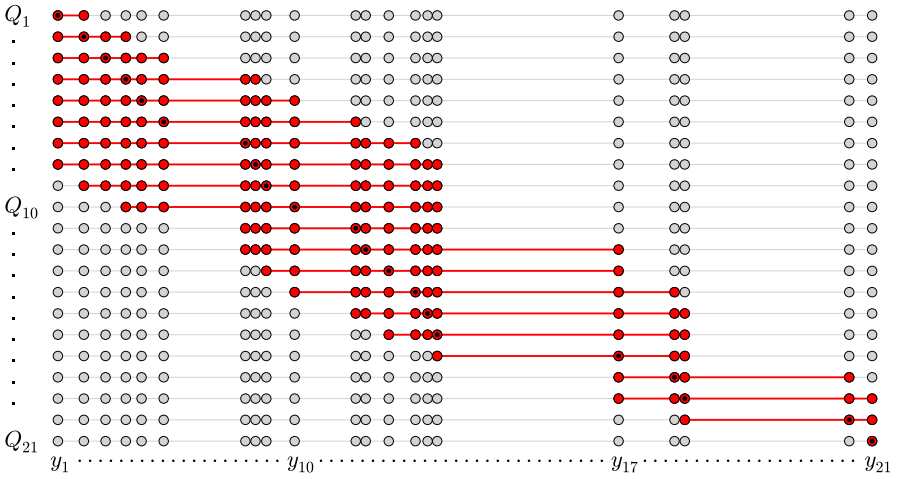


Fig. 2 Median sequence sets $Q_1, \ldots, Q_{21}$ (*colored in red* with median elements y_z marked by *black dot*) for 21 random values $y_1, \ldots, y_{21}$ w.r.t. objective function $f(Q) = |Q|/|P| - \texttt{smd}(Q)/\texttt{smd}(P)$—where $\texttt{smd}(\cdot)$ denotes sum of absolute deviations from the median; the sets are identical for any arbitrary dependence on the median $\texttt{med}(Q)$ that could be added to f, and for any such function the optimal value is attained among those 21 sets (Proposition 2) (Color figure online)

(e.g., $k^*_{10} = 13$, $k^*_{11} = 10$, and $k^*_{12} = 11$). The precise behavior of the k^*_z-sequence is determined by the cluster structure of the target values and the specific level-2 objective function. Below we will see that for some functions there is an additional regularity in the k^*_z-sequence that allows further algorithmic exploitation. For now, let us first note that, as desired, searching the median sequence is sufficient for finding optimal subsets of Q independent of the precise objective:

Proposition 2 *Let f be a dispersion-corrected objective function based on the median. Then the tight optimistic estimator of f can be computed as the maximum value on the median sequence, i.e., $\hat{f}(Q) = \max\{f(Q_z) \colon z \in [m]\}$.*

Proof For a sub-multiset $R \subseteq Q$ let us define the gap count $\gamma(R)$ as

$$\gamma(R) = |\{y \in Q \setminus R \colon \min R < y < \max R\}| \,.$$

Let $O \subseteq Q$ be an f-maximizer with minimal gap count, i.e., $f(R) < f(O)$ for all R with $\gamma(R) < \gamma(O)$. Assume that $\gamma(O) > 0$. That means there is a $y \in Q \setminus O$ such that $\min O < y < \max O$. Define

$$S = \begin{cases} (O \setminus \{\min O\}) \cup \{y\}, & \text{if } y \leq \mathtt{med}(O) \\ (O \setminus \{\max O\}) \cup \{y\}, & \text{otherwise} \end{cases} .$$

Per definition we have $|S| = |O|$ and $\mathtt{med}(S) = \mathtt{med}(O)$. Additionally, we can check that $S^{\mathtt{med}}_\Delta \leq_e O^{\mathtt{med}}_\Delta$, and, hence, $d(S) \leq d(Q)$. This implies that

$$f(S) = g(|S|, \mathtt{med}(S), d(S)) \geq g(|O|, \mathtt{med}(O), d(O)) = f(O) \,.$$

However, per definition of S it also holds that $\gamma(S) < \gamma(O)$, which contradicts that O is an f-optimizer with minimal gap count. Hence, any f-maximizer O must have a gap count of zero. In other words, O is of the form $O = Q^k_z$ as in Eq. (8) for some median $z \in [m]$ and some cardinality $k \in [m_z]$ and per definition we have $f(Q_z) \geq f(O)$ as required. □

Consequently, we can compute the tight optimistic estimator for any dispersion-corrected objective function based on the median in time $\mathcal{O}(m^2)$ for subpopulations of size m—again, given a suitable incremental formula for d. While this is not generally a practical algorithm in itself, it is a useful departure point for designing one. In the next section we show how it can be brought down to linear time when we introduce some additional constraints on the objective function.

3.3 Reaching linear time—objectives based on dispersion-corrected coverage

Equipped with the general concept of the median sequence, we can now address the special case of dispersion-corrected objective functions where the trade-off between the subpopulation size and target value dispersion is captured by a linear function of

size and the sum of absolute differences from the median. Concretely, let us define the **dispersion-corrected coverage** (w.r.t. absolute median deviation) by

$$\texttt{dcc}(Q) = \frac{|Q|}{|P|}\left(1 - \frac{\texttt{amd}(Q)}{\texttt{amd}(P)}\right)_+ = \left(\frac{|Q|}{|P|} - \frac{\texttt{smd}(Q)}{\texttt{smd}(P)}\right)_+$$

where $\texttt{smd}(Q) = \sum_{y \in Q} |y - \texttt{med}(Q)|$ denotes the **sum of absolute deviations from the median**. We then consider objective functions based on the dispersion-corrected coverage of the form

$$f(Q) = g(\texttt{dcc}(Q), \texttt{med}(Q)) \tag{9}$$

where g is non-decreasing in its first argument. Let us note, however, that we could replace the dcc function by any linear function that depends positively on $|Q|$ and negatively on smd. It is easy to verify that function of this form also obey the more general definition of level-2 objective functions given in Sec. 3.2, and, hence can be optimized via the median sequence.

The key to computing the tight optimistic estimator $\hat{f}$ in linear time for functions based on dispersion-corrected coverage is then that the members of the median sequence Q_z can be computed incrementally in constant time. Indeed, we can prove the following theorem, which states that the optimal size for a multiset around median index z is within 3 of the optimal size for a multiset around median index $z+1$—a fact that can also be observed in the example given in Fig. 2.

Theorem 3 *Let f be of the form of Eq. (9). For $z \in [m-1]$ it holds for the size k_z^* of the f-optimal multiset with median z that*

$$k_z^* \in \{\max(0, k_{z+1}^* - 3), \ldots, \min(m_z, k_{z+1}^* + 3)\}\,. \tag{10}$$

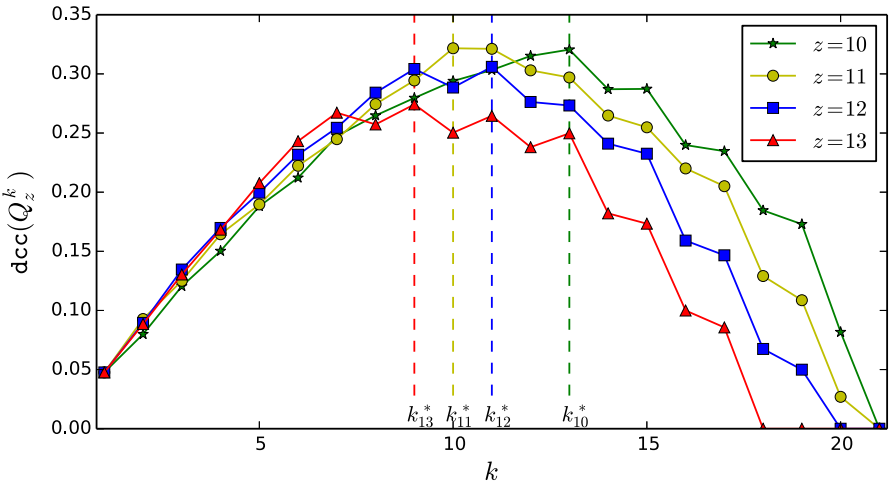


Fig. 3 Dispersion-corrected coverage of the sets Q_z^k as defined in Eq. (8) for median indices $z \in \{10, 11, 12, 13\}$ and the 21 random target values from Fig. 2; the sets Q_z can be found in incremental constant time since optimal size k_z^* is within a constant range of k_{z+1}^* (Theorem 3)

let Q be given by $\{y_1, \ldots, y_m\}$ in ascending order
compute $e_l(i)$ and $e_r(i)$ for $i \in [m]$ through Eqs. (11) and (12)
$f(Q_m) = g(1/|P|, y_m)$ and $k^*_m = 1$
for $z = m - 1$ **to** 1 **do**
  let $k^- = \max(0, k^*_{z+1} - 3)$ and $k^+ = \min(m_z, k^*_{z+1} + 3)$ with m_z as in Eq. (7)
  for $k = k^-$ **to** k^+ **do**
    let $a = z - \lfloor k/2 \rfloor$ and $b = z + \lceil k/2 \rceil$
    $\texttt{smd}(Q^k_z) = e_l(z) - e_l(a) - (a-1)(y_z - y_a) + e_r(z) - e_r(b) - (m-b)(y_b - y_z)$
    $f(Q^k_z) = g(k/|P| - \texttt{smd}(Q^k_z)/\texttt{smd}(P), y_z)$
  end
  $f(Q_z) = f(Q^{k^*_z}_z)$ with k^*_z s.t. $f(Q^{k^*_z}_z) = \max\{f(Q^k_z) \colon k^- \le k \le k^+\}$
end
$\hat{f}(Q) = \max\{f(Q_z) \colon z \in [m]\}$

Algorithm 2: Linear time algorithm for computing tight optimistic estimator $\hat{f}(Q)$ of objective $f(Q) = g(\texttt{dcc}(Q), \texttt{med}(Q))$ as in Eq. (9). After a linear time pre-processing to compute the cumulative left and right error terms, algorithm iterates over all possible median indices z—finding optimal value for median index $z+1$ in constant time based on optimal value for index z via Theorem 3 and Proposition 4.

One idea to prove this theorem is to show that (a) the gain in f for increasing the multiset around a median index z is alternating between two discrete concave functions and (b) that the gains for growing multisets between two consecutive median indices are bounding each other. For an intuitive understanding of this argument, Fig. 3 shows for four different median indices $z \in \{10, 11, 12, 13\}$ the dispersion-corrected coverage for the sets Q^k_z as a function in k. On closer inspection, we can observe that when considering only every second segment of each function graph, the corresponding dcc-values have a concave shape. A detailed proof, which is rather long and partially technical, can be found in "Appendix A".

It follows that, after computing the objective value of Q_m trivially as $f(Q_m) = g(1/|P|, y_m)$, we can obtain $f(Q_{z-1})$ for $z = m, \ldots, 2$ by checking the at most seven candidate set sizes given by Eq. (10) as

$$f(Q^{z-1}) = \max\left\{f(Q^{k^-_z}_{z-1}), \ldots, f(Q^{k^+_z}_{z-1})\right\}$$

with $k^-_z = \max(k^*_z - 3, 1)$ and $k^+_z = \min(k^*_z + 3, m_z)$. For this strategy to result in an overall linear time algorithm, it remains to see that we can compute individual evaluations of f in constant time (after some initial $\mathcal{O}(m)$ pre-processing step).

As a general data structure for quickly computing sums of absolute deviations from a center point, we can define for $i \in [m]$ the cumulative **left error** $e_l(i)$ and the cumulative **right error** $e_r(i)$ as

$$e_l(i) = \sum_{j=1}^{i-1} y_i - y_j, \qquad e_r(i) = \sum_{j=i+1}^{m} y_j - y_i\,.$$

Note that we can compute these error terms for all $i \in [m]$ by iterating over the ordered target values in time $\mathcal{O}(m)$ via the recursions

$$e_l(i) = e_l(i-1) + (i-1)(y_i - y_{i-1}) \tag{11}$$
$$e_r(i) = e_r(i+1) + (m-i)(y_{i+1} - y_i) \tag{12}$$

and $e_l(1) = e_r(m) = 0$. Subsequently, we can compute sums of deviations from center points of arbitrary subpopulations in constant time, as the following statement shows (see "Appendix B" for a proof).

Proposition 4 *Let $Q = \{y_1, \ldots, y_a, \ldots, y_z, \ldots, y_b, \ldots, y_m\}$ be a multiset with $1 \leq a < z < b \leq m$ and $y_i \leq y_j$ for $i \leq j$. Then the sum of absolute deviations to y_i of all elements of the submultiset $\{y_a, \ldots, y_z, \ldots, y_b\}$ can be expressed as*

$$\sum_{i=a}^{b} |y_z - y_i| = e_l(z) - e_l(a) - (a-1)(y_z - y_a) + e_r(z) - e_r(b) - (m-b)(y_b - y_z)\,.$$

With this we can compute $k \mapsto f(Q_z^k)$ in constant time (assuming g can be computed in constant time). Together with Proposition 2 and Theorem 3 this results in a linear time algorithm for computing $Q \mapsto \hat{f}(Q)$ (see Algorithm 2 for a pseudo-code that summarizes all ideas).

4 Dispersion-corrected subgroup discovery in practice

The overall result of Sect. 3 is an efficient algorithm for dispersion-corrected subgroup discovery which, e.g., allows us to replace the coverage term in standard objective functions by the dispersion-corrected coverage. To evaluate this efficiency claim as well as the value of dispersion-correction, let us consider as objective the normalized and dispersion-corrected impact function based on the median, i.e., $f_1(Q) = \mathtt{dcc}(Q)\mathtt{mds}_+(Q)$ where $\mathtt{mds}_+$ is the **positive relative median shift**

$$\mathtt{mds}_+(Q) = \left(\frac{\mathtt{med}(Q) - \mathtt{med}(P)}{\mathtt{max}(P) - \mathtt{med}(P)}\right)_+ .$$

This function obeys Eq. (9); thus, its tight optimistic estimator can be computed using the linear time algorithm from Sect. 3.3. The following empirical results were gathered by applying it to a range of publicly available real-world datasets.[4] We will first investigate the effect of dispersion-correction on the output before turning to the effect of the tight optimistic estimator on the computation time.

[4] Datasets contain all regression datasets from the KEEL repository (Alcalá et al. 2010) with at least 5 attributes and two materials datasets from the Nomad Repository nomad-coe.eu/; see Table. 1. Implementation available in open source Java library realKD bitbucket.org/realKD/. Computation times determined on MacBook Pro 3.1 GHz Intel Core i7.

4.1 Selection bias of dispersion-correction and its statistical merit

To investigate the selection bias of f_1 let us also consider the non-dispersion corrected variant $f_0(Q) = \texttt{cov}(Q)\texttt{mds}_+(Q)$, where we simply replace the dispersion-corrected coverage by the ordinary coverage. This function is a monotone level 1 function, hence, its tight optimistic estimator $\hat{f}_0$ can be computed in linear time using the top sequence approach. Figure 4 shows the characteristics of the optimal subgroups that are discovered with respect to both of these objective functions (see also Table. 1 for exact values) where for all datasets the language of closed conjunctions $\mathcal{C}_{\text{cnj}}$ has been used as description language.

The first observation is that—as enforced by design—for all datasets the mean absolute deviation from the median is lower for the dispersion-corrected variant (except in one case where both functions yield the same subgroup). On average the dispersion for f_1 is 49 percent of the global dispersion, whereas it is 113 percent for f_0, i.e., *when not optimizing the dispersion it is on average higher in the subgroups than in the global population*. When it comes to the other subgroup characteristics, coverage and median target value, the global picture is that f_1 discovers somewhat more specific groups (mean coverage 0.3 versus 0.44 for f_0) with higher median shift (on average 0.73 normalized median deviations higher). However, in contrast to dispersion, the behavior for median shift and coverage varies across the datasets. In Fig. 4, the datasets are ordered according to the difference in subgroup medians between the optimal subgroups w.r.t. f_0 and those w.r.t. f_1. This ordering reveals the following categorization of outcomes: When our description language is not able to reduce the error of subgroups with very high median value, f_1 settles for more coherent groups with a less extreme but still outstanding central tendency. On the other end of the scale,

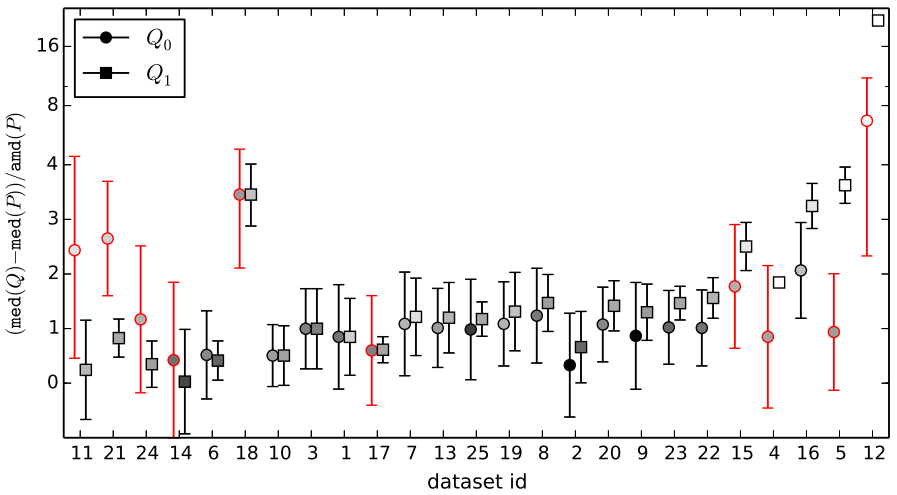


Fig. 4 Normalized median of optimal subgroup w.r.t. uncorrected positive median shift (Q_0) and w.r.t. dispersion-corrected positive median shift (Q_1) for 25 test datasets, sorted according to median difference. *Error bars* show mean absolute median deviation of subgroups; groups marked *red* have larger deviation than global deviation; *fill color* indicates group coverage from 0 (*white*) to 1 (*black*) (Color figure online)

Table 1 Datasets with corresponding population size ($|P|$), number of base propositions ($|\Pi|$), global median ($\texttt{med}(P)$) and mean absolute median deviation ($\texttt{amd}(P)$) followed by coverage ($\texttt{cov}(Q_0)$, $\texttt{cov}(Q_1)$), median ($\texttt{med}(Q_0)$, $\texttt{med}(Q_1)$), and mean absolute median deviation ($\texttt{amd}(Q_0)$, $\texttt{amd}(Q_1)$) for best subgroup w.r.t. non-dispersion corrected function f_0 and dispersion-corrected function f_1, respectively

Dataset							Selection Bias						Efficiency				
	Name	Target	$\lvert P\rvert$	$\lvert\Pi\rvert$	$\texttt{med}(P)$	$\texttt{amd}(P)$	$\texttt{cov}(Q_0)$	$\texttt{cov}(Q_1)$	$\texttt{med}(Q_0)$	$\texttt{med}(Q_1)$	$\texttt{amd}(Q_0)$	$\texttt{amd}(Q_1)$	a_{eff}	$\lvert\mathcal{E}_0\rvert$	$\lvert\mathcal{E}_1\rvert$	t_0	t_1
1	abalone	rings	4,177	69	9	2.359	**0.544**	0.191	11	11	2.257	**1.662**	1	848,258	690,177	**304**	339
2	ailerons	goal	13,750	357	−0.0008	0.000303	**0.906**	0.59	−0.0007	**−0.0006**	0.000288	**0.000198**	0.3	1,069,456	54,103	6542	**460**
3	autoMPG8	mpg	392	24	22.5	6.524	0.497	0.497	29	29	4.791	4.791	1	96	67	0.11	**0.09**
4	baseball	salary	337	24	740	954.386	**0.362**	0.003	1550	**2500**	1245.092	**0**	1	117	117	0.22	**0.21**
5	california	med. h. value	20,640	72	179,700	88,354	**0.385**	0.019	262,500	**500,001**	94261	**294,00**	0.4	1,368,662	65,707	2676	**368**
6	compactiv	usr	8192	202	89	9.661	0.464	**0.603**	**94**	93	7.8	**3.472**	0.5	2,458,105	59,053	5161	**208**
7	concrete	compr. strength	1030	70	34.4	13.427	**0.284**	0.1291	48.97	**50.7**	12.744	**9.512**	1	512,195	221,322	43.9	**35.8**
8	dee	consume	365	60	2.787	0.831	**0.523**	0.381	3.815	**4.008**	0.721	**0.434**	1	18,663	2653	2.05	**1.29**
9	delta_ail	sa	7,129	66	−0.0001	0.000231	**0.902**	0.392	0.0001	**0.0002**	0.000226	**0.000119**	1	45,194	2632	33.3	**6.11**
10	delta_elv	se	9517	66	0.001	0.00198	**0.384**	0.369	0.002	0.002	0.00112	**0.00108**	1	10145	1415	8.9	**4.01**
11	elevators	goal	16,599	155	0.02	0.00411	0.113	**0.283**	**0.03**	0.021	0.00813	**0.00373**	0.05	6,356,465	526,114	13,712	**2891**
12	forestfires	area	517	70	0.52	12.832	**0.01**	0.002	86.45	**278.53**	56.027	**0**	1	340,426	264,207	**23**	23.7
13	friedman	output	1200	48	14.651	4.234	**0.387**	0.294	18.934	**19.727**	3.065	**2.73**	1	19,209	2,489	3.23	**1.56**
14	house	price	22,784	160	33,200	28,456	0.56	**0.723**	**45,200**	34,000	40,576	**27,214**	0.002	1,221,696	114,566	7937	**1308**
15	laser	output	993	42	46	35.561	**0.32**	0.093	109	**135**	40.313	**15.662**	1	2008	815	0.96	**0.83**
16	mortgage	30 y. rate	1049	128	6.71	2.373	**0.256**	0.097	11.61	**14.41**	2.081	**0.98**	1	40,753	1270	11.6	**1.59**

Table 1 continued

Dataset							Selection Bias						Efficiency				
	Name	Target	$\|P\|$	$\|\Pi\|$	med(P)	amd(P)	cov(Q_0)	cov(Q_1)	med(Q_0)	med(Q_1)	amd(Q_0)	amd(Q_1)	a_{eff}	$\|\mathcal{E}_0\|$	$\|\mathcal{E}_1\|$	t_0	t_1
17	mv	y	40,768	79	−5.02086	8.509	**0.497**	0.349	0.076	**0.193**	8.541	**2.032**	1	6513	1017	31.9	**13.2**
18	pole	output	14,998	260	0	28.949	**0.40**	0.24	100	100	38.995	**16.692**	0.2	1,041,146	2966	2638	**15**
19	puma32h	thetadd6	8192	318	0.000261	0.023	**0.299**	0.244	0.026	**0.031**	0.018	**0.017**	0.4	3,141,046	5782	2648	**15.5**
20	stock	company10	950	80	46.625	5.47	**0.471**	0.337	52.5	**54.375**	3.741	**2.515**	1	85,692	1822	12.5	**1.56**
21	treasury	1 m. def. rate	1049	128	6.61	2.473	0.182	**0.339**	**13.16**	8.65	2.591	**0.863**	1	49,197	9247	14.8	**5.91**
22	wankara	mean temp.	321	87	47.7	12.753	**0.545**	0.296	60.6	**67.6**	8.873	**4.752**	1	191,053	4081	11.9	**1.24**
23	wizmir	mean temp.	1,461	82	60	12.622	**0.6**	0.349	72.9	**78.5**	8.527	**3.889**	1	177,768	1409	38.5	**1.48**
24	binaries	delta E	82	499	0.106	0.277	0.305	**0.378**	**0.43**	0.202	0.373	**0.118**	0.5	4,712,128	204	1200	**0.29**
25	gold	Evdw-Evdw0	12,200	250	0.131	0.088	**0.765**	0.34	0.217	**0.234**	0.081	**0.0278**	0.4	1,498,185	451	5650	**3.96**

bold-face indicates higher coverage, higher median, and lower dispersion; underlines indicate higher dispersion than in global population; final column segment contains accuracy parameter used in the efficiency study (a_{eff}) as well as number of expanded nodes ($|\mathcal{E}_0|$, $|\mathcal{E}_1|$) and computation time in seconds (t_0, t_1) for optimistic estimator based on top sequence $\hat{f}_0$ and tight optimistic estimator $\hat{f}_1$, respectively—in both cases when optimizing f_1; depth-limit is 10 for all datasets with $a < 1$, no depth-limit otherwise

when no coherent groups with moderate size and median shift can be identified, the dispersion-corrected objective selects very small groups with the most extreme target values. The majority of datasets obey the global trend of dispersion-correction leading to somewhat more specific subgroups with higher median that are, as intended, more coherent.

To determine based on these empirical observations whether we should generally favor dispersion correction, we have to specify an application context that specifies the relative importance of coverage, central tendency, and dispersion. For that let us consider the common statistical setting in which we do not observe the full global population P but instead subgroup discovery is performed only on an i.i.d. sample $P' \subseteq P$ yielding subpopulations $Q' = \sigma(P')$. While σ has been optimized w.r.t. the statistics on that sample Q' we are actually interested in the properties of the full subpopulation $Q = \sigma(P)$. For instance, a natural question is what is the minimal y-value that we expect to see in a random individual $q \in Q$ with high confidence. That is, we prefer subgroups with an as high as possible threshold l such that a random $q \in Q$ satisfies with probability[5] $1 - \delta$ that $y(q) \geq l$. This criterion gives rise to a natural trade-off between the three evaluation metrics through the **empirical Chebycheff inequality** (see Kabán 2012, Eq. (17)), according to which we can compute such a value as $\texttt{mean}(Q') - \epsilon(Q')$ where

$$\epsilon(Q') = \sqrt{\frac{(|Q'|^2 - 1)\texttt{var}(Q')}{|Q'|^2\delta - |Q'|}}$$

and $\texttt{var}(Y) = \sum_{y \in Y}(y - \texttt{mean}(Y))^2/(|Y| - 1)$ is the sample variance. Note that this expression is only defined for sample subpopulations with a size of at least $1/\delta$. For smaller subgroups our best guess for a threshold value would be the one derived from the global sample $\texttt{mean}(P') - \epsilon(P')$ (which we assume to be large enough to determine an ϵ-value). This gives rise to the following **standardized lower confidence bound score** $\tilde{l}$ that evaluates how much a subgroup improves over the global l value:

$$\tilde{l}(Q') = \left(\frac{l(Q') - l(P')}{\sqrt{\texttt{var}(P')}}\right)_+ \text{ where } l(Q') = \begin{cases} \texttt{mean}(Q') - \epsilon(Q'), & \text{if } \epsilon(Q') \text{ defined} \\ \texttt{mean}(P') - \epsilon(P'), & \text{otherwise} \end{cases}.$$

The plot on the left side of Fig. 5 shows the score values of the optimal subgroup w.r.t. to f_1 ($\tilde{l}_1$) and f_0 ($\tilde{l}_0$) using confidence parameter $\delta = 0.05$. Except for three exceptions (datasets 3,4, and 12), the subgroup resulting from f_1 provides a higher lower bound than those from the non-dispersion corrected variant f_0. That is, the data shows a strong advantage for dispersion correction when we are interested in selectors that mostly select individuals with a high target value from the underlying population P. In order to test the significance of these results, we can employ the **Bayesian sign-test** (Benavoli et al. 2014), a modern alternative to classic frequentist null hypothesis tests that avoids many of the well-known disadvantages of those (see Demšar 2008; Benavoli et al. 2016). With Bayesian hypothesis tests, we can directly

[5] The probability is w.r.t. to the distribution with which the sample $P' \subseteq P$ is drawn.

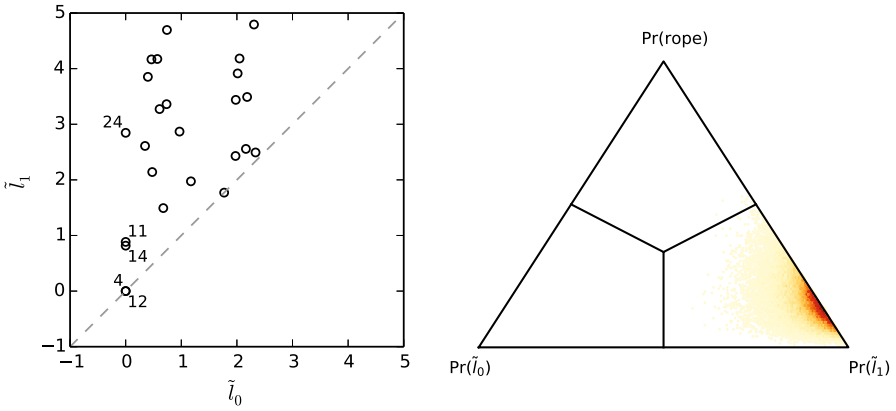


Fig. 5 Effect of dispersion correction on lower bound of 95-percent confidence interval of target variable; (*left*) improvement over global lower bound in standard deviations of dispersion-corrected objective ($\tilde{l}_1$) and non-dispersion-corrected objective ($\tilde{l}_0$) with annotations showing ids of datasets where either method provides no improvement; (*right*) posterior joint probabilities of the events that normalized difference $(\tilde{l}_1 - \tilde{l}_0)/\max\{\tilde{l}_0, \tilde{l}_1\}$ is larger than 0.1 (Pr($\tilde{l}_1$)), less than -0.1 (Pr($\tilde{l}_0$)), or within $[-0.1, 0.1]$ (Pr(rope)) according to Bayesian sign-test in barycentric coordinates (sections correspond to regions where corresponding event is maximum a posteriori outcome)

evaluate the posterior probabilities of hypotheses given our experimental data instead of just rejecting a null hypothesis based on some arbitrary significance level. Moreover, we differentiate between sample size and effect size by the introduction of a region of practical equivalence (rope). Here, we are interested in the relative difference $\tilde{z} = (\tilde{l}_1 - \tilde{l}_0)/(\max\{\tilde{l}_0, \tilde{l}_1\})$ on average for random subgroup discovery problems. Using a conservative choice for the rope, we call the two objective functions practically equivalent if the mean $\tilde{z}$-value is at most $r = 0.1$. Choosing the prior belief that f_0 is superior, i.e., $\tilde{z} < -r$, with a prior weight of 1, the procedure yields based on our 25 test datasets the posterior probability of approximately 1 that $\tilde{z} > r$ on average (see the right part of Fig. 5 for in illustration of the posterior belief). Hence, we can conclude that dispersion-correction improves the relative lower confidence bound of target values on average by more than 10 percent when compared to the non-dispersion-corrected function.

4.2 Efficiency of the tight optimistic estimator

To study the effect of the tight optimistic estimator, let us compare its performance to that of a baseline estimator that can be computed with the standard top sequence approach. Since f_1 is upper bounded by f_0, $\hat{f}_0$ is a valid, albeit non-tight, optimistic estimator for f_1 and can thus be used for this purpose. The exact speed-up factor is determined by the ratio of enumerated nodes for both variants as well as the ratio of computation times for an individual optimistic estimator computation. While both factors determine the practically relevant outcome, the number of nodes evaluated is a much more stable quantity, which indicates the full underlying speed-up potential independent of implementation details. Similarly, "number of nodes evaluated" is also

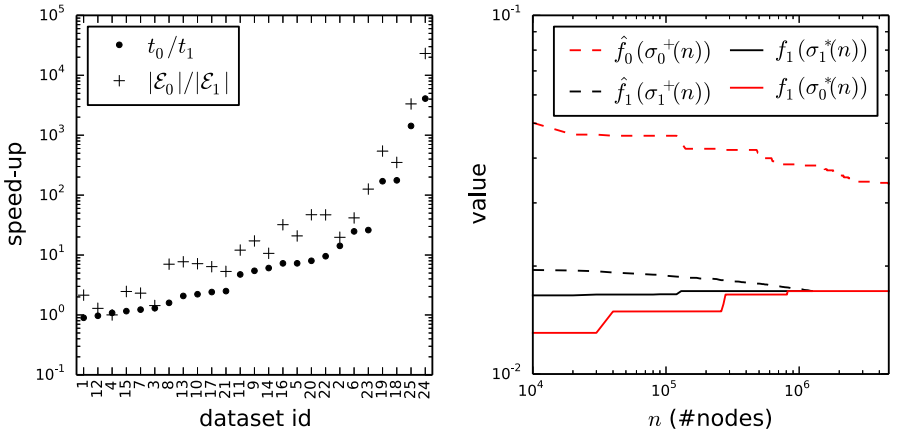


Fig. 6 Effect of tight optimistic estimator; (*right*) optimistic estimation ($\hat{f}_1(\sigma_1^+)$, $\hat{f}_0(\sigma_0^+)$) of remaining search space and value of current best solution ($f_1(\sigma_1^*)$, $f_1(\sigma_0^*)$) resulting from tight estimator and top sequence estimator, respectively—per processed nodes for dataset *binaries*; (*left*) speedup factor (t_0/t_1) in increasing order for all datasets plus node-reduction factor ($|\mathcal{E}_0|/|\mathcal{E}_1|$), which indicate potential of further improvements of estimator algorithm

an insightful unit of time for measuring optimization progress. Therefore, in addition to the computation time in seconds t_0 and t_1, let us denote by $\mathcal{E}_0, \mathcal{E}_1 \subseteq \mathcal{L}$ the set of nodes enumerated by branch-and-bound using $\hat{f}_0$ and $\hat{f}_1$, respectively—*but in both cases for optimizing the dispersion-corrected objective* f_1. Moreover, when running branch-and-bound with optimistic estimator $\hat{f}_i$, let us denote by $\sigma_i^*(n)$ and $\sigma_i^+(n)$ the best selector found and the top element of the priority queue (w.r.t. $\hat{f}_i$), respectively, after n nodes have been enumerated.

Figure 6 (left) shows the speed-up factor t_1/t_0 on a logarithmic axis for all datasets in increasing order along with the potential speed-up factors $|\mathcal{E}_0|/|\mathcal{E}_1|$ (see Table 1 for numerical values). There are seven datasets for which the speed-up is minor followed by four datasets with a modest speed-up factor of 2. For the remaining 14 datasets, however, we have substantial speed-up factors between 4 and 20 and in four cases immense values between 100 and 4000. This demonstrates the decisive potential effect of tight value estimation even when compared to another non-trivial estimator like $\hat{f}_0$ (which itself improves over simpler options by orders of magnitude; see Lemmerich et al. 2016). Similar to the results in Sect. 4.1, the Bayesian sign-test for the normalized difference $z = (t_1 - t_0)/\max\{t_1, t_0\}$ with the prior set to practical equivalence ($z \in [-0.1, 0.1]$) reveals that the posterior probability of $\hat{f}_1$ being superior to $\hat{f}_0$ is apx. 1. In almost all cases the potential speed-up given by the ratio of enumerated nodes is considerably higher than the actual speed-up, which shows that, despite the same asymptotic time complexity, an individual computation of the tight optimistic estimator is slower than the simpler top sequence based estimator—but also indicates that there is room for improvements in the implementation.

Examining the optimization progress over time for the *binaries* dataset, which exhibits the most extreme speed-up (right plot in Fig. 6), we can see that not only does the tight optimistic estimator close the gap between best current selector and current

highest potential selector much faster—thus creating the huge speed-up factor—but also that it causes better solutions to be found earlier. This is an important property when we want to use the algorithm as an *anytime algorithm*, i.e., when allowing the user to terminate computation preemptively, which is important in interactive data analysis systems. This is an advantage enabled specifically by using the tight optimistic estimator in conjunction with the best-first node expansion strategy.

5 Conclusion

During the preceding sections, we developed and evaluated an effective algorithm for simultaneously optimizing size, central tendency, and dispersion in subgroup discovery with a numerical target. This algorithm is based on two central results: (1) the tight optimistic estimator for any objective function that is based on some dispersion measure around the median can be computed as the function's maximum on a linear-sized sequence of sets—the median sequence (Proposition 2); and (2) for objective functions based on the concept of the dispersion-corrected coverage w.r.t. the absolute deviation from the median, the individual sets of the median sequence can be generated in incremental constant time (Theorem 3).

Among the possible applications of the proposed approach, the perhaps most important one is to replace the standard coverage term in classic objective functions by the dispersion-corrected coverage, i.e., the relative subgroup size minus the relative subgroup dispersion, to reduce the error of result subgroups—where error refers to the descriptive or predictive inaccuracy incurred when assuming the median value of a subgroup for all its members. As we saw empirically for the impact function (based on the median), this correction also has a statistical advantage resulting in subgroups where we can assume larger target values for unseen group members with high confidence. In addition to enabling dispersion-correction to known objective functions, the presented algorithm also provides novel degrees of freedom, which might be interesting to exploit in their own right: The dependence on the median is not required to be monotone, which allows to incorporate a more sophisticated influence of the central tendency value than simple monotone average shifts. For instance, given a suitable statistical model for the global distribution, the effect of the median could be a function of the probability $\mathbb{P}[\mathtt{med}(Q)]$, e.g., its Shannon information content. Furthermore, the feasible dispersion measures allow for interesting weighting schemes, which include possibilities of asymmetric effects of the error (e.g., for only punishing one-sided deviation from the median). More generally, let us note that numerical subgroup discovery algorithms are also often applicable in settings where numerical association rules are sought (see Aumann and Lindell 2003). The appeal of branch-and-bound optimization is here that it circumvents the expensive enumeration step of all frequent (high coverage) sets.

Regarding the limitations of the presented approach, let us note that it cannot be directly applied to the previously proposed dispersion-aware functions, i.e., the t-score $\mathtt{tsc}(Q) = \sqrt{|Q|}(\mathtt{mean}(Q) - \mathtt{mean}(P))/\mathtt{std}(Q)$ and the mmad score for ranked data $\mathtt{mmd}(Q) = |Q|/(2\mathtt{med}(Q) + \mathtt{mmd}(Q))$. While both of these functions can be optimized via the median sequence approach (assuming a t-score variant based on the

median), we are lacking an efficient incremental formula for computing the individual function values for all median sequence sets, i.e., a replacement for Theorem 3. Though finding such a replacement in future research is conceivable, this leaves us for the moment with a quadratic time algorithm (in the subgroup size) for the tight optimistic estimator, which is not generally feasible (although potentially useful for smaller datasets or as part of a hybrid optimistic estimator, which uses the approach for sufficiently small subgroups only).

Since they share basic monotonicities, it is possible to use functions based on dispersion-corrected coverage as an optimization proxy for the above mentioned objectives. For instance, the ranking of the top 20 subgroups w.r.t. the dispersion-corrected binomial quality function, $\mathtt{dcb}(Q) = \sqrt{\mathtt{dcc}(Q)}(\mathtt{med}(Q) - \mathtt{med}(P))$, turns out to have a mean Spearman rank correlation coefficient with the median-based t-score of apx. 0.783 on five randomly selected test datasets (*delta_elv*, *laser*, *stock*, *treasury*, *gold*). However, a more systematic understanding of the differences and commonalities of these functions is necessary to reliably replace them with one another. Moreover, the correlation deteriorates quite sharply when we compare to the original mean/variance based t-score (mean Spearman correlation coefficient 0.567), which points to the perhaps more fundamental limitation of the presented approach for dispersion-correction: it relies on using the median as measure of central tendency. While the median and the mean absolute deviation from the median are an interpretable, robust, and sound combination of measures (the median of a set of values minimizes the sum of absolute deviations), the mean and the variance are just as sound, are potentially more relevant when sensitivity to outliers is required, and provide a wealth of statistical tools (e.g., Chebyshev's inequality used above).

Hence, a straightforward but valuable direction for future work is the extension of efficient tight optimistic estimator computation to dispersion-correction based on the mean and variance. A basic observation for this task is that objective functions based on dispersion measures around the mean must also attain their maximum on gap-free intervals of target values. However, for a given collection of target values, there is a quadratic number of intervals such that a further idea is required in order to attain an efficient, i.e., (log-)linear time algorithm. Another valuable direction for future research is the extension of consistency and error optimization to the case of multidimensional target variables where subgroup parameters can represent complex statistical models (known as *exceptional model mining* Duivesteijn et al. 2016). While this setting is algorithmically more challenging than the univariate case covered here, the underlying motivation remains: balancing group size and exceptionality, i.e., distance of local to global model parameters, with consistency, i.e., local model fit, should lead to the discovery of more meaningful statements about the data and the underlying domain.

Acknowledgements Open access funding provided by Max Planck Society. The authors thank the anonymous reviewers for their useful and constructive suggestions. Jilles Vreeken and Mario Boley are supported by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government. Bryan R. Goldsmith acknowledges support from the Alexander von Humboldt-Foundation with a Postdoctoral Fellowship. Additionally, this work was supported through the European Union's Horizon 2020 research and innovation program under Grant agreement No. 676580 with The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence.

Appendix A: Proof of Theorem 3

In order to proof Theorem 3, let us start by noting that for functions of the form of Eq. (9), finding the set size k_z^* corresponds to maximizing the dispersion-corrected coverage among all multisets with consecutive elements around median y_z (as defined in Eq. 8). In order to analyze this problem, let us write

$$h_z(k) = \texttt{dcc}(Q_z^k) = \frac{|Q_z^k|}{|P|} - \frac{\texttt{smd}(Q_z^k)}{\texttt{smd}(P)}$$

for the dispersion-corrected coverage of the multiset Q_z^k. Let $\Delta h_z \colon [m_z] \to \mathbb{R}$ denote the difference or gain function of h_z, i.e., $\Delta h_z(k) = h_z(k) - h_z(k-1)$ where we consider $Q_z^0 = \emptyset$ and, hence, $h_z(0) = 0$. With this definition we can show that h_z is alternating between two concave functions, i.e., considering either only the even or only the odd subset of its domain, the gains are monotonically decreasing. More precisely:

Lemma 5 *For all $k \in [m_z] \setminus \{1, 2\}$ we have that $\Delta h_z(k) \leq \Delta h_z(k-2)$.*

Proof For $k \in [m_z]$, let us denote by q_z^k the additional y-value that Q_z^k contains compared to Q_z^{k-1} (considering $Q_z^0 = \emptyset$), i.e., $Q_z^k \setminus Q_z^{k-1} = \{q_z^k\}$. We can check that

$$q_z^k = \begin{cases} q_{z-\left\lfloor \frac{k-1}{2} \right\rfloor}, & k \text{ odd} \\ q_{z+\left\lceil \frac{k-1}{2} \right\rceil}, & k \text{ even} \end{cases}.$$

With this and using the shorthands $n = |P|$ and $d = \texttt{smd}(P)$ we can write

$$\begin{aligned}
\Delta h_z(k) - \Delta h_z(k-2) &= h_z(k) - h_z(k-1) - (h_z(k-2) - h_z(k-3)) \\
&= \frac{k}{n} - \frac{\texttt{smd}(Q_z^k)}{d} - \frac{k-1}{n} + \frac{\texttt{smd}(Q_z^{k-1})}{d} - \frac{k-2}{n} \\
&\quad + \frac{\texttt{smd}(Q_z^{k-2})}{d} + \frac{k-3}{n} - \frac{\texttt{smd}(Q_z^{k-3})}{d} \\
&= \frac{1}{n} \underbrace{(k - k + 1 - k + 2 + k - 3)}_{=0} + \frac{1}{d} \Big(\texttt{smd}(Q_z^{k-2}) \\
&\quad - \texttt{smd}(Q_z^k) + \texttt{smd}(Q_z^{k-1}) - \texttt{smd}(Q_z^{k-3}) \Big) \\
&= \frac{1}{d} \Big(-|q_z^k - y_z| - |q_z^{k-1} - y_z| + |q_z^{k-1} - y_z|
\end{aligned}$$

$$
\begin{aligned}
&\quad + |q_z^{k-2} - y_z|\Big) \\
&= \frac{1}{d}\left(-|q_z^k - y_z| + |q_z^{k-2} - y_z|\right) \\
&\text{case } k \text{ odd} \\
&= \frac{1}{d}\left(-\left(y_z - y_{z-\left\lfloor \frac{k-1}{2} \right\rfloor}\right) + \left(y_z - y_{z-\left\lfloor \frac{k-3}{2} \right\rfloor}\right)\right) \\
&= y_{z-\left\lfloor \frac{k-1}{2} \right\rfloor} - y_{z-\left\lfloor \frac{k-3}{2} \right\rfloor} \leq 0 \\
&\text{case } k \text{ even} \\
&= \frac{1}{d}\left(-\left(y_{z+\left\lceil \frac{k-1}{2} \right\rceil} - y_z\right) + \left(y_{z+\left\lceil \frac{k-3}{2} \right\rceil} - y_z\right)\right) \\
&= y_{z+\left\lceil \frac{k-3}{2} \right\rceil} - y_{z+\left\lceil \frac{k-1}{2} \right\rceil} \leq 0
\end{aligned}
$$

□

One important consequence of this fact is that the operation of growing a set around median z by two elements—one to the left and one to the right—has monotonically decreasing gains. In other words, the smoothed function $h_z(k) = h_z(k) + h_z(k-1)$ is concave or formally

$$\Delta h_z(k) + \Delta h_z(k-1) \geq \Delta h_z(k+1) + \Delta h_z(k)\,. \tag{13}$$

Moreover, we can relate the gain functions of consecutive median indices as follows.

Lemma 6 *Let $z \in [m] \setminus \{1\}$ and $k \in [m_{z-1}] \setminus \{1, 2, 3\}$. It holds that*

$$\Delta h_{z-1}(k-2) + \Delta h_{z-1}(k-3) \geq \Delta h_z(k) + \Delta h_z(k-1) \tag{14}$$

$$\Delta h_{z-1}(k) + \Delta h_{z-1}(k-1) \leq \Delta h_z(k-2) + \Delta h_z(k-3) \tag{15}$$

Proof For this proof, let us use the same shorthands as in the proof of Lemma 5 and start by noting that for all $i \in [m]$ and $k \in [m_z] \setminus \{1\}$ we have the equality

$$\Delta h_i(k) + \Delta h_i(k-1) = \frac{2}{n} - \frac{|q_i^k - y_i| + |q_i^{k-1} - y_i|}{d} \tag{16}$$

which we can see by extending

$$
\begin{aligned}
\Delta h_i(k) + \Delta h_i(k-1) &= h_i(k) - h_i(k-1) + h_i(k-1) - h_i(k-2) \\
&= \frac{k-k+2}{n} - \frac{\mathtt{smd}(Q_i^k) - \mathtt{smd}(Q_i^{k-2})}{d} \\
&= \frac{2}{n} - \frac{|q_i^k - y_i| + |q_i^{k-1} - y_i|}{d}\,.
\end{aligned}
$$

We can then show Eq. (14) by applying Eq. (16) two times to

$$\begin{aligned}
&\Delta h_{z-1}(k-2) + \Delta h_{z-1}(k-3) - (\Delta h_z(k) + \Delta h_z(k-1)) \\
&\quad = \frac{1}{d}\left(-|q_{z-1}^{k-2} - y_{z-1}| - |q_{z-1}^{k-3} - y_{z-1}| + |q_z^k - y_z| + |q_z^{k-1} - y_z|\right)
\end{aligned}$$

and finally by checking separately the case k odd

$$\begin{aligned}
&= \frac{1}{d}\left(y_{z-1-\left\lfloor\frac{k-3}{2}\right\rfloor} - y_{z-1} + y_{z-1} - y_{z-1+\left\lceil\frac{k-4}{2}\right\rceil} + y_z - y_{z-\left\lfloor\frac{k-1}{2}\right\rfloor} + y_{z+\left\lceil\frac{k-2}{2}\right\rceil} - y_z\right) \\
&= \frac{1}{d}\left(\underbrace{y_{z-\left\lfloor\frac{k-1}{2}\right\rfloor} - y_{z-\left\lfloor\frac{k-1}{2}\right\rfloor}}_{=0} + \underbrace{y_{z-1+\left\lceil\frac{k}{2}\right\rceil} - y_{z-3+\left\lceil\frac{k}{2}\right\rceil}}_{\geq 0}\right) \geq 0
\end{aligned}$$

and the case k even

$$\begin{aligned}
&= \frac{1}{d}\left(y_{z-1} - y_{z-1+\left\lceil\frac{k-3}{2}\right\rceil} + y_{z-1-\left\lfloor\frac{k-4}{2}\right\rfloor} - y_{z-1} + y_{z+\left\lceil\frac{k-1}{2}\right\rceil} - y_z + y_z - y_{z-\left\lfloor\frac{k-2}{2}\right\rfloor}\right) \\
&= \frac{1}{d}\left(\underbrace{y_{z+1-\left\lfloor\frac{k}{2}\right\rfloor} - y_{z+1-\left\lfloor\frac{k}{2}\right\rfloor}}_{=0} + \underbrace{y_{z+\left\lceil\frac{k-1}{2}\right\rceil} - y_{z-2+\left\lceil\frac{k-1}{2}\right\rceil}}_{\geq 0}\right) \geq 0\,.
\end{aligned}$$

Similarly, for Eq. (15) by applying Eq. (16) two times we can write

$$\begin{aligned}
&\Delta h_{z-1}(k) + \Delta h_{z-1}(k-1) - (\Delta h_z(k-2) + \Delta h_z(k-3)) \\
&\quad = \frac{1}{d}\left(-|q_{z-1}^{k} - y_{z-1}| - |q_{z-1}^{k-1} - y_{z-1}| + |q_z^{k-2} - y_z| + |q_z^{k-3} - y_z|\right) \\
&\quad = \begin{cases}
\frac{1}{d}\left(\underbrace{y_{z-1-\left\lfloor\frac{k-1}{2}\right\rfloor} - y_{z+1-\left\lfloor\frac{k-1}{2}\right\rfloor}}_{\leq 0} + \underbrace{y_{z-2+\left\lceil\frac{k}{2}\right\rceil} - y_{z-2+\left\lceil\frac{k}{2}\right\rceil}}_{=0}\right) \leq 0, & k \text{ odd} \\
\frac{1}{d}\left(\underbrace{y_{z-\left\lfloor\frac{k}{2}\right\rfloor} - y_{z+2-\left\lfloor\frac{k}{2}\right\rfloor}}_{\leq 0} + \underbrace{y_{z-1+\left\lceil\frac{k-1}{2}\right\rceil} - y_{z-1+\left\lceil\frac{k-1}{2}\right\rceil}}_{=0}\right) \leq 0, & k \text{ even}
\end{cases}
\end{aligned}$$

□

Combining all of the above we can finally proof our main result as follows.

Proof (Theorem 3) We start by showing that every $k \in [m_{z+1}]$ with $k < k_z^* - 3$ can not be an optimizer of h_{z+1}. It follows that $k_z^* - 3 \leq k_{z+1}^*$, and, hence, $k_z^* \leq k_{z+1}^* + 3$ as required for the upper bound. Indeed, we have

$$\begin{aligned} h_{z+1}(k) &= h_{z+1}(k+2) - (\Delta h_{z+1}(k+2) + \Delta h_{z+1}(k+1)) \\ &\leq h_{z+1}(k+2) - (\Delta h_{z+1}(k_z^* - 2) + \Delta h_{z+1}(k_z^* - 3)) && \text{(by Eq. (13))} \\ &\leq h_{z+1}(k+2) - \underbrace{(\Delta h_z(k_z^*) + \Delta h_z(k_z^* - 1))}_{>0 \text{ by def. of } k_z^*} < h_{z+1}(k+2)\,. && \text{(by Lm. 6)} \end{aligned}$$

Analogously, for the lower bound, we show that every $k \in [m_{z+1}]$ with $k > k_z^* + 3$ can not be the smallest optimizer of h_{z+1}. It follows that $k_z^* + 3 \geq k_{z+1}^*$, and, hence, $k_z^* \geq k_{z+1}^* - 3$ as required. Indeed, we can write

$$\begin{aligned} h_{z+1}(k) &= h_{z+1}(k-2) + \Delta h_{z+1}(k) + \Delta h_{k+1}(k-1) \\ &\leq h_{z+1}(k-2) + \Delta h_{z+1}(k_z^* + 4) + \Delta h_{z+1}(k_z^* + 3) && \text{(by Eq. (13))} \\ &\leq h_{z+1}(k-2) + \underbrace{\Delta h_z(k_z^* + 2) + \Delta h_z(k_z^* + 1)}_{\leq 0 \text{ by def. of } k_z^*} \leq h_{z+1}(k-2) && \text{(by Lm. 6)} \end{aligned}$$

□

Appendix B: Additional proofs

Proof (Proposition 4) Using d_{ij} as a shorthand for $y_j - y_i$ for $i, j \in [m]$ with $i \leq j$ we can write

$$\begin{aligned} & e_l(z) - e_l(a) - (a-1)(y_z - y_a) + e_r(z) - e_r(b) - (m-b)d_{zb} \\ &= \sum_{i=1}^{z-1} d_{iz} - \sum_{i=1}^{a-1} d_{ia} - (a-1)d_{az} + \sum_{i=z+1}^{m} d_{zi} - \sum_{i=b+1}^{m} d_{bi} - (m-b)d_{zb} \\ &= \sum_{i=a}^{z-1} d_{iz} + \sum_{i=1}^{a-1} \underbrace{(d_{iz} - d_{ia})}_{d_{az}} - (a-1)d_{az} + \sum_{i=z+1}^{b} d_{zi} \\ &\quad + \sum_{i=b+1}^{m} \underbrace{(d_{zi} - d_{bi})}_{d_{zb}} - (m-b)d_{zb} \\ &= \sum_{i=a}^{z-1} d_{iz} + (a-1)d_{az} - (a-1)d_{az} + \sum_{i=z+1}^{b} d_{zi} + (m-b)d_{zb} - (m-b)d_{zb} \\ &= \sum_{i=a}^{z-1} d_{iz} + \sum_{i=z+1}^{b} d_{zi} = \sum_{i=a}^{b} |y_z - y_i| \end{aligned}$$

□

Appendix C: Summary of used notations

Symbol	Meaning	Defined in
$\lvert\cdot\rvert$	Cardinality of a set or absolute value of a number	–
$[k]$	Set of integers $\{1, \ldots, k\}$	2
$(x)_+$	$\max\{x, 0\}$ for a real-valued expression x	2
$\leq_e$	Element-wise less-or-equal relation for multisets of real values	3
2^X	Power set of a set X, i.e., set of all of its subsets	–
$\mathbb{N}^X$	Set of all multisets containing elements from set X	–
σ, φ	Subgroup selectors $\sigma, \varphi\colon P \to \{\text{true}, \text{false}\}$	2.1
c	Measure of central tendency	3.1
d	Measure of dispersion	3.2
$e_l(i), e_r(i)$	Left and right cumulative errors of target values up to value i	3.3
f	Objective function	2.1
$\hat{f}$	Tight optimistic estimator of objective function f	2.2
m	Number of elements in subpopulation Q	3
m_z	Maximal size parameter k for consecutive value set Q_z^k	3.2
k_z^*	f-maximizing size parameter k for consecutive value set Q_z^k	3.2
y	Numeric target attribute $y\colon P \to \mathbb{R}$	2.1
y_i	i-th target value of subpopulation w.r.t. ascending order	3
P	Global population of given subgroup discovery problem	2.1
Q	Some subpopulation $Q \subseteq P$	2.1
Q_z	Median sequence element with median index z	3.2
Q_z^k	Submultiset of Q with k consecutive elements around index z	3.2
T_i	Top sequence element i, i.e., $T_i = \{y_{m-i+1}, \ldots, y_m\}$	3.1
Y	Real-valued multiset	3
$Y_\Delta^{\mathtt{med}}$	multiset of differences of elements in Y to its median	3.2
$\mathcal{L}, \mathcal{L}_{\text{cnj}}$	Description language and language of conjunctions	2.1
$\mathcal{C}_{\text{cnj}}$	Language of closed conjunctions	2.1
$\mathtt{amd}(Q)$	Mean absolute deviation of y-values in Q to their median	3.2
$\mathtt{cov}(Q)$	Coverage, i.e., relative size $\lvert Q\rvert/\lvert P\rvert$ of subpopulation Q	2.1
$\mathtt{dcc}(Q)$	Dispersion-corrected coverage of subpopulation Q	3.3
$\mathtt{mean}(Q)$	Arithmetic mean of y-values in Q	–
$\mathtt{ipa}(Q)$	Impact, i.e., weighted mean-shift, of subpopulation Q	2.1
$\mathtt{med}(Q)$	Median of y-values in Q	3.1
$\mathtt{smd}(Q)$	Sum of absolute deviations of y-values in Q to their median	3.3

References

Alcalá J, Fernández A, Luengo J, Derrac J, García S, Sánchez L, Herrera F (2010) Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Mult Valued Log Soft Comput 17(2–3):255–287

Atzmueller M (2015) Subgroup discovery. Wiley Interdiscip Rev Data Min Knowl Discov 5(1):35–49

Aumann Y, Lindell Y (2003) A statistical theory for quantitative association rules. J Intell Inf Syst 20(3):255–283

Bay SD, Pazzani MJ (2001) Detecting group differences: mining contrast sets. Data Min Knowl Discov 5(3):213–246

Benavoli A, Corani G, Mangili F, Zaffalon M, Ruggeri F (2014) A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In: ICML. pp 1026–1034

Benavoli A, Corani G, Demsar J, Zaffalon M (2016) Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. arXiv:1606.04316

Boley M, Grosskreutz H (2009) Non-redundant subgroup discovery using a closure system. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 179–194
Boley M, Moens S, Gärtner T (2012) Linear space direct pattern sampling using coupling from the past. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 69–77
Demšar J (2008) On the appropriateness of statistical tests in machine learning. In: Workshop on evaluation methods for machine learning in conjunction with ICML
Duivesteijn W, Knobbe A (2011) Exploiting false discoveries-statistical validation of patterns and quality measures in subgroup discovery. IEEE 11th international conference on data mining. IEEE, pp 151–160
Duivesteijn W, Feelders AJ, Knobbe A (2016) Exceptional model mining. Data Min Knowl Discov 30(1):47–98
Friedman JH, Fisher NI (1999) Bump hunting in high-dimensional data. Stat Comput 9(2):123–143
Goldsmith BR, Boley M, Vreeken J, Scheffler M, Ghiringhelli LM (2017) Uncovering structure-property relationships of materials by subgroup discovery. New J Phys 19(1):13–31
Grosskreutz H, Rüping S, Wrobel S (2008) Tight optimistic estimates for fast subgroup discovery. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 440–456
Grosskreutz H, Boley M, Krause-Traudes M (2010) Subgroup discovery for election analysis: a case study in descriptive data mining. In: International conference on discovery science. Springer, pp 57–71
Huan J, Wang W, Prins J (2003) Efficient mining of frequent subgraphs in the presence of isomorphism. In: 3rd IEEE international conference on data mining. IEEE, pp 549–552
Kabán A (2012) Non-parametric detection of meaningless distances in high dimensional data. Stat Comput 22(2):375–385
Klösgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, pp 249–271
Klösgen W (2002) Data mining tasks and methods: subgroup discovery: deviation analysis. In: Handbook of data mining and knowledge discovery. Oxford University Press Inc., pp 354–361
Lavrač N, Kavšek B, Flach P, Todorovski L (2004) Subgroup discovery with cn2-sd. J Mach Learn Res 5:153–188
Lemmerich F, Atzmueller M, Puppe F (2016) Fast exhaustive subgroup discovery with numerical target concepts. Data Min Knowl Discov 30(3):711–762
Li G, Zaki MJ (2016) Sampling frequent and minimal boolean patterns: theory and application in classification. Data Min Knowl Discov 30(1):181–225
Mehlhorn K, Sanders P (2008) Algorithms and data structures: the basic toolbox. Springer, Berlin
Parthasarathy S, Zaki MJ, Ogihara M, Dwarkadas S (1999) Incremental and interactive sequence mining. In: Proceedings of 8th international conference on information and knowledge management. ACM, pp 251–258
Pasquier N, Bastide Y, Taouil R, Lakhal L (1999) Efficient mining of association rules using closed itemset lattices. Inf Syst 24(1):25–46
Pieters BF, Knobbe A, Dzeroski S (2010) Subgroup discovery in ranked data, with an application to gene set enrichment. In: Proceedings preference learning workshop (PL 2010) at ECML PKDD, vol 10. pp 1–18
Schmidt J, Hapfelmeier A, Mueller M, Perneczky R, Kurz A, Drzezga A, Kramer S (2010) Interpreting pet scans by structured patient data: a data mining case study in dementia research. Knowl Inf Syst 24(1):149–170
Song H, Kull M, Flach P, Kalogridis G (2016) Subgroup discovery with proper scoring rules. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 492–510
Uno T, Asai T, Uchida Y, Arimura H (2004) An efficient algorithm for enumerating closed patterns in transaction databases. In: International conference on discovery science. Springer, pp 16–31
Webb GI (1995) Opus: an efficient admissible algorithm for unordered search. J Artif Intell Res 3:431–465
Webb GI (2001) Discovering associations with numeric variables. In: Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 383–388
Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: European symposium on principles of data mining and knowledge discovery. Springer, pp 78–87