

# Efficiently Discovering Locally Exceptional yet Globally Representative Subgroups

Janis Kalofolias, Mario Boley, and Jilles Vreeken  
Max Planck Institute for Informatics and Saarland University  
Saarland Informatics Campus, Germany  
{kalofolias,mboley,jilles}@mpi-inf.mpg.de

**Abstract**—Subgroup discovery is a local pattern mining technique to find interpretable descriptions of sub-populations that stand out on a given target variable. That is, these sub-populations are exceptional with regard to the global distribution. In this paper we argue that in many applications, such as scientific discovery, subgroups are only useful if they are additionally representative of the global distribution with regard to a control variable. That is, when the distribution of this control variable is the same, or almost the same, as over the whole data.

We formalise this objective function and give an efficient algorithm to compute its tight optimistic estimator for the case of a numeric target and a binary control variable. This enables us to use the branch-and-bound framework to efficiently discover the top- $k$  subgroups that are both exceptional as well as representative. Experimental evaluation on a wide range of datasets shows that with this algorithm we discover meaningful representative patterns and are up to orders of magnitude faster in terms of node evaluations as well as time.

**Index Terms**—Subgroup discovery, Branch-and-bound, Fairness

## I. INTRODUCTION

Pattern mining in general, and subgroup discovery in particular, are powerful exploratory data mining techniques that can reveal important local structure that can easily be missed, or not explicitly represented, by global models [2]. More precisely, subgroup discovery aims to find interpretable selectors of local data regions by optimising a trade-off between exceptionality, i.e., the degree to which the distribution of a designated target variable varies locally from its global distribution, and generality, i.e., the fraction of the data space covered by the selector.

A problem with this traditional approach is its simplistic notion of generality: if a subpopulation is relatively sizeable it is considered general, even though it might show arbitrary statistical obscurities. This lack of representativeness is a key problem in many important scenarios.

In scientific discovery and theory development, we often seek to identify local factors that influence some variable, but want to control for the influence of other potential explanations. For instance, in materials science we may want to discover structural patterns that characterise the HOMO-LUMO energy gap in gold nanoclusters [11], independent of the parity of their atom count, which is already known to have a strong influence. As another example, in political science we are often interested in discovering demographics with a high

affinity to a certain political party. However, findings should not rediscover known geographic influences (See Fig. 1).

Besides science, there are other examples where traditional subgroup discovery fails. In policy development and other fairness-aware applications there are often ethical and legal requirements that demand the distribution of policy recipients to match the underlying population w.r.t. to some sensitive variable. For instance, while students with a high chance of obtaining a degree are reasonable candidates for defining the application criteria of a scholarship, we might still want to ensure that the eligible population is gender-balanced.

All of the above settings share the requirement of subgroups to not only be relatively sizeable, but also statistically *representative* w.r.t. some control variable. Specifically, this variable should have a similar distribution between the subgroup and the global population, exhibiting what is called **statistical parity** [26]. In contrast, simply removing the control variable, to avoid it influencing the result, is infeasible, since it can usually be approximately recovered by the remaining variables (known as *red-lining effect* [5]). See again Fig. 1.

While there are several techniques to enforce representativeness of binary global classifiers, it is unclear how those can (i) be generalised to settings that go beyond a binary prediction task, and (ii) be integrated effectively into branch-and-bound, the standard framework for optimal subgroup discovery. This framework requires an efficiently, i.e., near linear time, computable optimistic estimator for the desired objective function.

Therefore, in this work we propose a general representativeness term that can be incorporated into subgroup discovery objective functions, which is based on the statistical distance between the local and the global distribution of the control variable. Moreover, we show how the resulting representative subgroup discovery problem can be solved efficiently for the case of a binary control and a numeric target variable. In particular, we propose RAWR, an algorithm to compute the tight optimistic estimator for the representativeness-aware objective function, in  $O(n \log n)$  time. Experiments show that, when employing this algorithm in the branch-and-bound framework, we can prune orders of magnitude of candidates in comparison to the state of the art, which, besides reducing memory consumption, leads to orders of magnitude gain in runtime; therewith, RAWR makes it possible to mine representative subgroups in otherwise computationally infeasible settings.

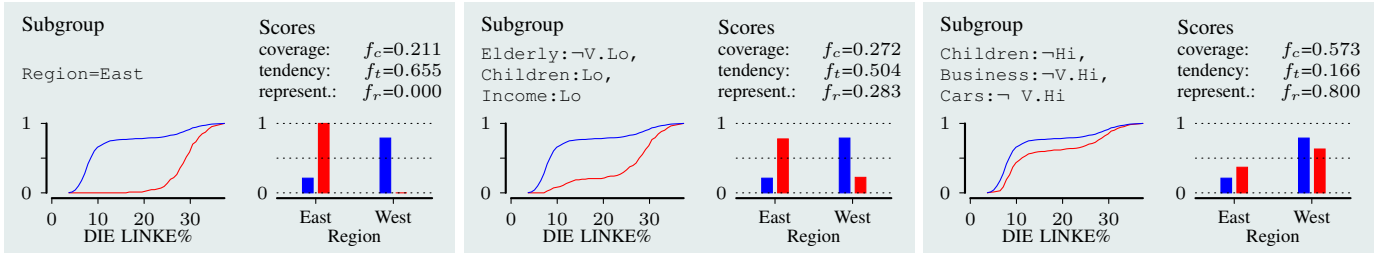


Figure 1: Subgroups of German voting districts in 2009 elections with high percentage of the left-wing party “DIE LINKE”. Blocks show the cdf of percentage in the subgroup (red) and global population (blue), along with the distribution of district locations. Traditional subgroup discovery (left) recovers the main trend: eastern districts support “DIE LINKE”. Removing the *region* attribute (middle) results in a similar subgroup. Only when explicitly controlling for geography (right) do we discover subgroups that stand out with regard to voting behaviour, while at the same time being representative for the whole country.

## II. PRELIMINARIES

In this section we recall the necessary basics of subgroup discovery and Branch and Bound (B&B) search.

### A. Subgroup discovery

The goal in subgroup discovery is to identify useful subpopulations of a given **global population**  $P$ , which can be viewed as a set of  $n$  entities  $P = \{e_1, \dots, e_n\}$ . These subpopulations are identified by Boolean functions,  $\sigma : P \rightarrow \{\top, \perp\}$ , the **subgroup selectors**, each of which defines a subpopulation  $Q = \text{ext}(\sigma)$  through the extension function  $\text{ext} : \sigma \mapsto \{e \in P : \sigma(e) = \top\}$ ; note that we will often use  $\sigma$  and  $Q$  interchangeably. The set of all available subgroup selectors, the **selection language**  $\mathcal{L}$ , most commonly comprises conjunctions formulated over a set of basic descriptive conditions, e.g.,  $[\text{age} > 18]$  or  $[\text{sex} = \text{‘Male’}]$ . In this paper, however, it suffices to consider an abstract selection language.

Additionally, we assume a continuous **target variable**  $y : P \rightarrow \mathbb{R}$  and a discrete **control variable**  $c : P \rightarrow \{1, \dots, K\}$ . The usefulness of a subgroup can then be encoded by a real-valued **objective function**  $f : 2^P \rightarrow \mathbb{R}$ . An exemplary such function, for numeric target variables, is the **impact function**

$$f_{\text{ct}}(Q) := f_c(Q)f_t(Q) = \frac{|Q|}{|P|} \frac{\bar{y}_Q - \bar{y}_P}{\max_{e \in P} y(e) - \bar{y}_P}, \quad (1)$$

where  $\bar{y}_P := \text{mean}\{y(e) : e \in P\}$  is the mean of the target values in the population, and  $\bar{y}_Q$  is the mean of those in  $Q$ .

A subgroup  $Q$  with a high  $f_{\text{ct}}$  value is exceptional, as the **central tendency factor**

$$f_t(Q) := \frac{\bar{y}_Q - \bar{y}_P}{\max_{e \in P} y(e) - \bar{y}_P}$$

ensures a high mean deviation of  $y$  within the subgroup. At the same time,  $Q$  exhibits a basic notion of generality, provided by the **coverage factor**  $f_c(Q) := |Q|/|P|$ .

In Sec. III we will augment this objective function to also represent a statistical notion of generality of  $Q$  w.r.t. the control variable.

### B. Branch and Bound with Optimistic Estimators

The standard algorithm for finding a set of  $k$  optimal subgroup selectors is Branch and Bound (here we give a basic overview and refer to Boley et al. [3] for more details). This algorithm employs a refinement operator  $\rho : \mathcal{L} \rightarrow 2^{\mathcal{L}}$ , as well as an **optimistic estimator**  $\hat{f}$  of the objective function  $f$ ,

$$\hat{f}(Q) \geq \max_{R \subseteq Q, R \neq \emptyset} f(R), \quad \forall Q \subseteq P. \quad (2)$$

The algorithm maintains a priority queue of candidate subgroup selectors  $\sigma \in \mathcal{L}$ , initialised to contain only a root selector, describing the entire population  $P$ . While keeping track of the subgroup  $Q^*$  with the the best value discovered so far, the algorithm iterates by picking from the queue that selector (resp. subgroup)  $Q = \text{ext}(\sigma)$  with the highest optimistic estimator value  $\hat{f}(Q)$ ; this favours subgroups with the greatest potential for improvement, resulting in a best-first-search scheme. If the optimistic estimator of  $Q$  ensures that none of its subgroups surpass the current best, i.e.,  $\hat{f}(Q) < f(Q^*)$ , all its refinements  $\rho(\sigma)$  can be safely pruned; otherwise, these refinements are placed in the queue. This procedure iterates until the queue empties, which guarantees to find the best, or by an easy extension, the best  $k$  subgroups.

Obviously, as the bound of the optimistic estimator gets tighter, the pruning potential increases, to become maximal when Eq. (2) holds with equality. Then we refer to  $\hat{f}$  as a **tight optimistic estimator** [12] of the objective function  $f$ .

Optionally, and to achieve even better pruning, the B&B algorithm may use the relaxed comparison  $\alpha \hat{f}(Q) > f(Q^*)$ , for an approximation factor  $\alpha \in (0, 1]$ , where a value of  $\alpha = 1$  yields the best subgroup. Lower  $\alpha$  generally yield better pruning, while guaranteeing that the discovered subgroup has a value no less than  $\alpha$  times that of the best subgroup.

The impact function  $f_{\text{ct}}$  allows an efficient implementation of its tight optimistic estimator, (since  $\max_{Q \subseteq P} f_{\text{ct}}(Q) = f_{\text{ct}}(Q^*)$ , with  $Q^* = \{e : y(e) \geq \bar{y}_P\}$ ), computable in linear time [3], [19]. We refer to this implementation as Binary Representativeness IGnorant (BRIG), to remind that  $f_{\text{ct}}$  is oblivious to the control variable.

In the next section we develop RAWR, an efficient algorithm to compute the tight optimistic estimator for the controlled

impact function, and also show that any optimistic estimator of the impact function (and thus also BRIG) can be used as a non-tight stand-in for the augmented one.

### III. REPRESENTATIVE SUBGROUP DISCOVERY

In order to describe the theoretical contributions of this paper, let us fix the following notation. We consider subpopulations  $Q \subseteq P$ , whose items we assume to be ordered in decreasing target value. Hence,  $y_i$  is the item of  $Q$  with the  $i$ -th greatest target value, which has a control class of  $c_i$ . Out of those elements of  $Q$  with class  $k$ , we denote  $y_i^{(k)}$  the one with the  $i$ -th greatest target value, and by  $n_k(Q) := |\{e \in Q : c(e) = k\}|$  their count, which we also refer to as the  $k$ -th **class count**. Similarly, we define the **class probability vector**  $\mathbf{p}(Q)$  with elements the **class probabilities**  $p_k(Q) := n_k(Q)/|Q|$ , for each class  $k$ .

#### A. The controlled impact function

We now augment the standard objective function of Eq (1) to also account for a broader notion of generality than coverage: the statistical generality of the subgroup w.r.t. the control variable. Specifically, we add a **representativeness factor**  $f_r(Q)$ , quantifying the similarity of the control distribution between  $Q$  and  $P$ . This forms the **controlled impact function**

$$f(Q) := f_{\text{ct}}(Q)^{1-\gamma} f_r(Q)^\gamma, \quad (3)$$

where the  $\gamma \in [0, 1)$  parameter tunes the trade-off between representativeness and the typical properties quantified by  $f_{\text{ct}}$ .

Viewed probabilistically, our goal is to select subpopulations independently of the control variable. This is equivalent to requiring, for a random entity  $e \in P$  from the population, that

$$\mathbb{P}(c(e)|\sigma(e) = \top) = \mathbb{P}(c(e)) \iff d(\mathbf{q}, \mathbf{p}) = 0,$$

where  $d$  is some distance measure between distributions with  $\mathbf{q} := \mathbf{p}(Q)$  and  $\mathbf{p} := \mathbf{p}(P)$ . In this work, we further fix  $d$  to be the **total variation distance**  $d(\mathbf{q}, \mathbf{p}) = \frac{1}{2} \sum_k |q_k - p_k|$ , equal to the maximal difference between probabilities of any set of control classes. This measure is at once intuitively interpretable and simple enough to allow efficient calculations.

Similar to the coverage and the central tendency factors, we design this new factor to assume values in the interval  $f_r \in [0, 1]$ , with more representative subgroups scoring higher:

$$f_r(Q) := 1 - \frac{d(\mathbf{p}, \mathbf{q}) - d_{\max}}{d_{\max}}, \quad d_{\max} := \max_{R \subseteq P} d(\mathbf{p}, \mathbf{r}).$$

We note that an important consequence of these bounds is that any optimistic estimator for the impact function is also a valid, albeit non-tight, optimistic estimator for the controlled impact function. This corresponds to fixing the added factor to its maximal value of  $f_r = 1$ .

Having introduced all constituents of the controlled impact function, we now proceed with the computation of its tight optimistic estimator. We first introduce a transform of the domain of the original optimisation problem from exponential to polynomial size in Sec. III-B. We then employ this transform in Sec. III-C to derive an efficient algorithm that computes this

tight optimistic estimator in  $O(n \log n)$  time, for the special case of a population with balanced binary classes.

#### B. Searching in the class counting space (CCS)

In this section we describe a transformation which aggregates the exponentially many subsets of  $Q$  in the original optimisation problem of Eq. (2) into polynomially many sets of subsets. Additionally, the maximum  $f$  value attained by any subset within each of these sets can be efficiently computed. From now on, we call any subset  $R \subseteq Q$  a **refinement** of  $Q$ .

For any given subgroup  $Q$ , we consider the space of all possible class count vectors  $\mathbf{I} := (n_1(Q), \dots, n_K(Q))$  that any refinement  $R \subseteq Q$  might assume,

$$\mathcal{I}(Q) := \prod_{k=1}^K \{0, \dots, n_k(Q)\}.$$

This space, which we refer to as the **class counting space** (CCS), is a subset of the lattice  $\mathbb{Z}^K$ , and partitions the original space  $2^Q$  into  $|\mathcal{I}(Q)| = \prod_{k=1}^K (n_k(Q) + 1)$  partitions. Each of these partitions, called the **equi-count refinement sets**  $\mathcal{R}_{\mathbf{I}}(Q)$ , consists of these refinements of  $Q$  with  $I_k$  items of control class  $k$ , for each class  $k = 1, \dots, K$ ,

$$\mathcal{R}_{\mathbf{I}}(Q) := \{R \subseteq Q : n_k(R) = I_k, \quad k = 1, \dots, K\}.$$

For an example of a CCS with  $K = 2$  classes see Fig. 2.

The computation of the tight optimistic estimator  $\hat{f}(Q) := \max_{R \in 2^Q} f(R)$  of Eq (2) can now be expressed as

$$\hat{f}(Q) := \max_{\mathbf{I} \in \mathcal{I}(Q)} \max_{R \in \mathcal{R}_{\mathbf{I}}(Q)} f(R) = \max_{\mathbf{I} \in \mathcal{I}(Q)} f^Q(\mathbf{I}),$$

where  $f^Q(\mathbf{I})$  refers to the maximal value attained over all refinements in the equi-count refinement set  $\mathcal{R}_{\mathbf{I}}$

$$f^Q(\mathbf{I}) := \begin{cases} \max_{R \in \mathcal{R}_{\mathbf{I}}(Q)} f(R) & \mathbf{I} \in \mathcal{I}(Q) \setminus \{\mathbf{0}\} \\ -\infty & \mathbf{I} = \mathbf{0} \end{cases}.$$

Similarly, the maxima of the impact function, central tendency and representativeness values over all refinements within  $\mathcal{R}_{\mathbf{I}}$  are denoted  $f_{\text{ct}}^Q(\mathbf{I})$ ,  $f_t^Q(\mathbf{I})$  and  $f_r^Q(\mathbf{I})$ , respectively.

In the next proposition we derive a closed form for the optimiser of  $f(Q)$  within an equi-count refinement set  $\mathcal{R}_{\mathbf{I}}$ , which can then be used to compute  $f_{\text{ct}}^Q(\mathbf{I})$  and thus  $f^Q(\mathbf{I})$ .

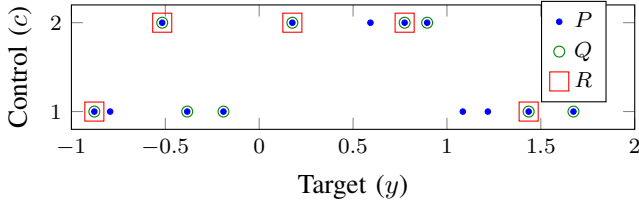
**Proposition 1.** *The optimal value  $f_{\text{ct}}^Q(\mathbf{I})$  is attained by the set*

$$R_{\mathbf{I}}^* := \bigcup_{k=1}^K \{y_1^{(k)}, \dots, y_{I_k}^{(k)}\},$$

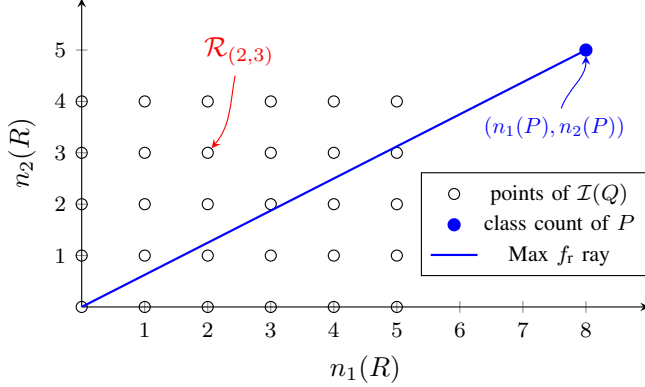
*which contains the  $I_k$  items with the greatest target value among those with control class  $k$ , for all classes  $k = 1, \dots, K$ .*

*Proof.* All sets  $R \in \mathcal{R}_{\mathbf{I}}(Q)$  have a constant coverage  $|R| = \sum_{k=1}^K I_k$ , so that maximising the objective value is equivalent to maximising the central tendency factor  $f_t^Q$ . We will show that  $R_{\mathbf{I}}^*$  attains the greatest  $f_t$  value over  $\mathcal{R}_{\mathbf{I}}(Q)$  by contradiction.

Assume there is a refinement  $R' \in \mathcal{R}_{\mathbf{I}}$  with  $R' \neq R_{\mathbf{I}}^*$  and  $f_t(R') > f_t(R_{\mathbf{I}}^*)$ . Since  $R_{\mathbf{I}}^*$  contains the items with maximum



(a) Toy population  $P$ , consisting of  $n_1(P) = 8$  items of control class 1 and  $n_2(P) = 5$  of class 2, a subgroup  $Q \subseteq P$ , with  $n_1(Q) = 5$  items of class 1 and  $n_2(Q) = 4$  of class 2, and a refinement  $R \subset Q$ .



(b) The CCS of  $Q$ , denoted  $\mathcal{I}(Q)$ , and the maximum  $f_r^Q$  ray. Each refinement  $R \subseteq Q$  with class counts  $\mathbf{I} = (I_1, I_2)$ , where  $I_1 = n_1(R)$  and  $I_2 = n_2(R)$ , is contained in the equi-class refinement set  $\mathcal{R}_{\mathbf{I}}$ , which corresponds to the point  $\mathbf{I}$  in  $\mathcal{I}(Q) = \{0, \dots, 5\} \times \{0, \dots, 4\}$ . Points closer to the  $\max f_r$  ray have a class count probability (ratio) closer to that of  $P$  and thus a higher  $f_r$  score.

Figure 2: The class counting space (bottom) for a toy population with  $K = 2$  control classes (top). The refinement  $R$  is contained in  $\mathcal{R}_{(2,3)}$ , corresponding to the annotated point.

$y$  value for each class, there is at least one sequence of refinements  $(R^{(0)}, \dots, R^{(T)})$ , starting with  $R^{(0)} := R_{\mathbf{I}}^*$  and ending at  $R^{(T)} := R^l$ , so that at each index  $\tau$  we exchange a single element between  $R^\tau$  with another in  $Q \setminus R^\tau$  of the same class, but a smaller target value. Formally,  $R^{(\tau)} = (R^{(\tau-1)} \setminus \{e\}) \cup \{e'\}$ , such that  $c(e') = c(e)$  and  $y(e') < y(e)$ . This implies that, for each  $\tau = 2, \dots, T$  we get

$$\sum_{e \in R^{(\tau)}} y(e) - \sum_{e \in R^{(\tau-1)}} y(e) = y(e') - y(e) < 0.$$

Dividing these sums with  $\sum_{k=1}^K I_k$ , turns them into means, and since  $f_t$  is increasing w.r.t. the target value mean, we have  $f_t(R^{(\tau)}) < f_t(R^{(\tau-1)})$ . By transitivity, it is  $f_t(R^l) < f_t(R_{\mathbf{I}}^*)$ , contradicting the optimality of  $f_t(R^l)$ .  $\square$

As a result, we can express the target value mean of the optimiser  $R_{\mathbf{I}}^*$  as  $\text{mean}(R_{\mathbf{I}}^*) = \sum_{k=1}^K \sum_{i=1}^{I_k} y_i^{(k)} / \|\mathbf{I}\|_1$ , where  $\|\mathbf{I}\|_1 = \sum_{k=1}^K I_k$  is the cardinality of each refinement in  $\mathcal{R}_{\mathbf{I}}$ . Now the impact function  $f_{\text{ct}}$  of Eq. (1) can be transformed onto the CCS as

$$f_{\text{ct}}^Q(\mathbf{I}) := \alpha_t \sum_{k=1}^K \sum_{i=1}^{I_k} y_i^{(k)} - \alpha_c \|\mathbf{I}\|_1, \quad (4)$$

where

$$\alpha_t = \frac{1}{\nu} > 0, \quad \alpha_c = \frac{\bar{y}_P}{\nu}, \quad \text{and} \quad \nu = |P| \left( \max_{e \in P} y(e) - \bar{y}_P \right).$$

Since the representativeness factor  $f_r(Q)$  depends only on the class counts of  $Q$ , it remains constant over  $\mathcal{R}_{\mathbf{I}}$  and does not affect the maximiser. Therefore, the transformed controlled impact function can be written as

$$f^Q(\mathbf{I}) := f_{\text{ct}}^Q(\mathbf{I})^{1-\gamma} \cdot f_r^Q(\mathbf{I})^\gamma \quad \gamma \in [0, 1].$$

Notice that the value  $f_{\text{ct}}^Q(\mathbf{I})$  can be computed in constant time for any point  $\mathbf{I} \in \mathcal{I}(Q)$ , after a pre-processing step of linear time. Indeed, assuming the items of a candidate subgroup are in decreasing order of target values, we can achieve this by first passing through the values and creating  $K$  cumulative sums of target values, one for each class; after this step, the value  $f_r^Q(\mathbf{I})$  can be easily retrieved as the sum of indices  $I_k$  within the cumulative sum for each class  $k$ , appropriately scaled. The controlled impact function  $f^Q(\mathbf{I})$  can be computed with trivial extra work to compute  $f_r^Q(\mathbf{I})$ .

Therefore, this transform can be used in a straightforward way to derive an algorithm to compute the tight optimistic estimator in  $O(n^K)$  time. However, a practical algorithm can benefit from further improvement, achieved in the next section for a special case of a population.

### C. A linearithmic algorithm for balanced binary controls

We now present a linearithmic algorithm able to compute the tight optimistic estimator of the controlled impact function of Eq. (3) for the case of a population  $P$  with balanced binary control classes, i.e.,  $c: P \rightarrow \{1, 2\}$  with  $n_1(P) = n_2(P)$ .

The rest of the analysis can be summarised in two key steps. First, we show that there is a sub-region of  $\mathcal{I}(Q)$  where  $f^Q(\mathbf{I})$  attains its maximum and then we present an efficient algorithm to search within this sub-region.

For this purpose we study the two factors,  $f_{\text{ct}}^Q$  and  $f_r^Q$  within the CCS. Both these factors form sequences that exhibit an appropriate notion of convexity for sequences, borrowed mutatis mutandis from Yucer [25]: a sequence  $a: N \rightarrow \mathbb{R}$ , with  $N = \{0, \dots, n\}$  and  $n \leq \infty$ , is called a **convex sequence** over  $N$ , if for all  $x, y \in N$  and each  $\lambda \in (0, 1)$

$$\lambda a^{(x)} + (1 - \lambda)a^{(y)} \geq \min_{u \in [z]} a^{(u)}, \quad z = \lambda x + (1 - \lambda)y.$$

Further, we call  $a$  a **concave sequence** if  $-a$  is convex.

We now study the  $f_{\text{ct}}^Q$  values, as  $|Q| = I_1 + I_2$  increases.

**Definition 1** (Optimal c-t path on  $\mathcal{I}(Q)$ ). *Let  $\pi^{(\mu)} \in \mathcal{I}(Q)$  be the maximiser of the  $f_{\text{ct}}^Q$  value among all points in the CCS with a fixed sum  $\mu$*

$$\pi^{(\mu)} := \arg \max_{\mathbf{I} \in \mathcal{I}, \|\mathbf{I}\|_1 = \mu} f_{\text{ct}}^Q(\mathbf{I}), \quad 0 \leq \mu \leq |Q|. \quad (5)$$

*We refer to the optimal point sequence  $\pi = (\pi^{(0)}, \dots, \pi^{(|Q|)})$  as the **optimal c-t path**.*

The optimal c-t path exhibits useful properties, discussed in the following lemma (for the proof see Appendix A).

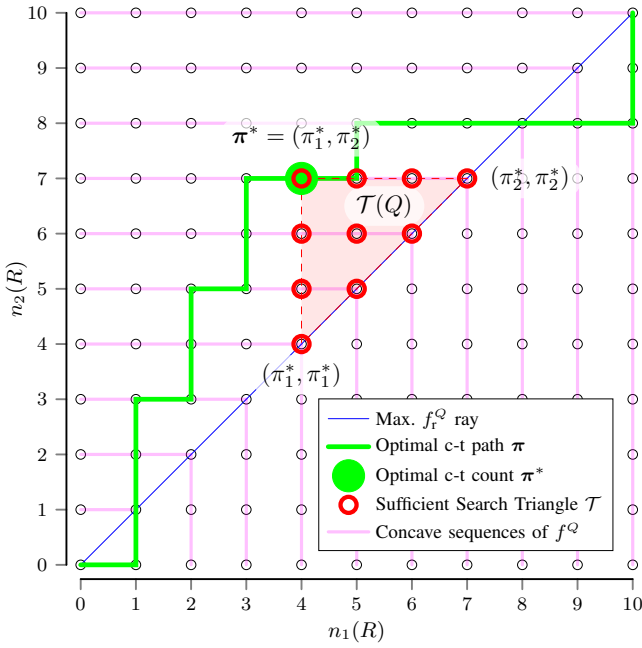


Figure 3: sufficient search triangle  $\mathcal{T}$  (red circles) in the CCS  $\mathcal{I}$ , and optimal point  $\pi^*$  along the c-t path (crooked line), which defines the 3 vertices of  $\mathcal{T}(Q)$ . We seek  $\hat{f} = \max f^Q$ , which lies on  $\mathcal{T}(Q)$ , and ternary search finds it efficiently along the concave sequences of  $f^Q$  (vertical/horizontal lines).

**Lemma 2.** Let  $e_1 = (1, 0)^T$  and  $e_2 = (0, 1)^T$  be the standard basis vectors of  $\mathbb{R}^2$ . Then (i) the  $\mu$ -th element of the optimal c-t path is the class count of the first  $\mu$  elements of  $P$ ; formally,

$$\pi^{(\mu)} = \sum_{i=1}^{\mu} e_{c_i} \quad 0 < \mu \leq |Q| \quad \text{and} \quad \pi^{(\mu)} = \mathbf{0}. \quad (6)$$

Moreover, (ii) the sequence  $f_{\text{ct}}^Q \circ \pi$ , with elements the  $f_{\text{ct}}^Q$  values computed along the c-t path  $\pi$ , is a concave sequence.

Lemma 2 allows for an  $O(\log n)$  algorithm to find the **optimal c-t point**  $\pi^* := \arg \max_{\mathbf{I} \in \mathcal{I}(Q)} f_{\text{ct}}^Q(\mathbf{I})$ , as we call the point  $\mathbf{f}$  the CCS that maximises the  $f_{\text{ct}}^Q$  value. Indeed,

$$f_{\text{ct}}(\pi^*) = \max_{0 \leq \mu \leq |Q|} \max_{\mathbf{I} \in \mathcal{I}(Q), |\mathbf{I}|=\mu} f_{\text{ct}}(\mathbf{I}) = \max_{0 \leq \mu \leq |Q|} f_{\text{ct}}(\pi^{(\mu)}),$$

where the last maximum runs over the  $f_{\text{ct}}^Q$  values of the optimal path sequence. Due to the concavity of this sequence, its maximum can be computed in  $O(\log n)$  time, using for instance the **ternary search** algorithm.

We now study the representativeness factor  $f_r(Q)$ , whose transform on the CCS for balanced binary controls becomes

$$f_r^Q(\mathbf{I}) := 1 - \left| 1 - \frac{2I_1}{I_1 + I_2} \right| = 1 - \left| 1 - \frac{2I_2}{I_1 + I_2} \right|.$$

We observe that the subgroups  $R \in Q$  that maximise this factor must have the same control class distribution as the population. Therefore, these subgroups must have an equal control class count  $n_1(R) = n_2(R)$ , and thus belong to those equi-count refinement sets  $\mathcal{R}_{\mathbf{I}}$ , for which  $I_1 = I_2$ . These, in

turn, lie on the so-called **maximum  $f_r^Q$  ray**  $\mathbf{I} = (a, a)^T$ ,  $a \geq 0$ . One example of a maximum  $f_r^Q$  ray appears in Fig. 3.

We now state a key theoretical proposition of this section, showing that it suffices to search for the optimal solution on a specific triangle of the CCS (for the proof see Appendix B).

**Proposition 3.** The maximum of the controlled impact function  $f^Q$  is attained at a point which lies in the (filled) triangle  $\mathcal{T}(Q) := \{(\pi_1^*, \pi_1^*), (\pi_2^*, \pi_2^*), \pi^*\}$ , with vertices the optimal c-t point  $\pi^* = (\pi_1^*, \pi_2^*)$  and its horizontal and vertical projections onto the maximum  $f_r^Q$  ray. We call this region the **sufficient search triangle**.

The sufficiency of the SST reduces the search space by at least half, which happens in the worst case scenario that the optimal c-t point  $\pi^*$  is on the North-West or South-East points. More importantly, we can efficiently optimise  $f^Q$  along specific directions within this region.

We now describe these directions. For each ordinate  $i_2 \in 0, \dots, n_2(Q)$  let the (West-to-East) **horizontal sequence** be

$$\mathbf{h}_{i_2} := (\mathbf{h}_{i_2}^{(0)}, \dots, \mathbf{h}_{i_2}^{(n_1(Q))}) = ((0, i_2), \dots, (n_1(Q), i_2)).$$

Similarly, for each abscissa  $i_1 \in 0, \dots, n_1(Q)$  we define the (South-to-North) **vertical sequence**

$$\mathbf{v}_{i_1} := (\mathbf{v}_{i_1}^{(0)}, \dots, \mathbf{v}_{i_1}^{(n_2(Q))}) = ((i_1, 0), \dots, (i_1, n_2(Q))).$$

When the transformed controlled impact function  $f^Q(\mathbf{I})$  is computed along the elements of certain of those sequences, it forms concave sequences, as we show below (for the proof see Appendix C), with the direct implication that the maximal value of  $f$  along these sequences can be computed in  $O(\log n)$ .

**Proposition 4.** Consider the values of the controlled impact function  $f^Q$  as they are computed along a horizontal sequence  $\mathbf{h}_{i_2}$ ; these form the sequence  $(f^Q \circ \mathbf{h}_{i_2})(\mu)$ , which for  $\mu \leq i_2$  is a concave sequence preceding the maximum  $f_r^Q$  ray. Similarly,  $(f^Q \circ \mathbf{v}_{i_1})(\mu)$  is a concave sequence for  $\mu \leq i_1$ .

Observing the example of the concave sequences of  $f^Q$  along the horizontal and vertical directions shown in Fig. 3, we notice that we can cover the entire SST with an appropriate selection of these concave sequences. This allows for an efficient optimisation procedure requiring  $O(n \log n)$  time, which is described in Algorithm 1 and operates as follows.

First, the optimal c-t point  $\pi^*$  is computed in  $O(\log n)$  time, along the concave sequence  $\pi$  (line 1); this point locates the SST. If  $\pi^*$  lies above the maximum  $f_r^Q$  ray (line 3-7), the points of  $\mathcal{T}(Q)$  lie along horizontal sub-sequences preceding the maximum  $f_r^Q$  ray; the  $f^Q$  values along each of them form a concave sequence, whose maximum can be found in  $O(\log n)$  (ln. 6). There are at most  $n_2(Q)$  such directions in  $\mathcal{T}(Q)$ , and they can all be scanned (ln. 5-7) in a total of  $O(n \log n)$  time. Similarly, when  $\pi^*$  lies below the maximum  $f_r^Q$  line (ln. 8), we optimise along the vertical sub-sequences (ln. 9-12).

#### IV. RELATED WORK

Whereas subgroup discovery [17] is well-studied in general [2], [4], [8], [12], [19], [24], to the best of our knowledge

---

**Algorithm 1: Representativeness AWare algoRithm**

---

**Input:** Population  $P$  (sorted w.r.t  $y$ , descending)  
**Input:** Subgroup  $Q$   
**Output:** Tight optimistic estimator  $\hat{f}$  of Eq. (3)

```
1  $\pi^* \leftarrow \mathbf{TernarySearch}$ (on  $f_{ct}^Q \circ \pi$  from 1 to  $|Q|$ );
2  $i_{beg} \leftarrow \min\{\pi_1^*, \pi_2^*\}$ ;
3 if  $\pi_1^* < \pi_2^*$  then
4    $i_{end} \leftarrow \min\{\pi_2^*, n_2(Q)\}$ ;
5   for  $i$  from  $i_{beg}$  to  $i_{end}$  do
6      $\phi \leftarrow \mathbf{TernarySearch}$ (on  $f^Q \circ h_i$  from  $i_{beg}$  to  $i_{end}$ );
7      $\hat{f} \leftarrow \max\{\hat{f}, \phi\}$ ;
8 else
9    $i_{end} \leftarrow \min\{\pi_1^*, n_1(Q)\}$ ;
10  for  $i$  from  $i_{beg}$  to  $i_{end}$  do
11     $\phi \leftarrow \mathbf{TernarySearch}$ (on  $f^Q \circ v_i$  from  $i_{beg}$  to  $i_{end}$ );
12     $\hat{f} \leftarrow \max\{\hat{f}, \phi\}$ ;
13 return  $\hat{f}$ ;
```

---

the notion of representativeness beyond coverage has not been studied in depth.

In pattern mining in general, there has been ample attention to iteratively discovering patterns that are surprising given background knowledge, for example with regard to permutation testing [10], [13], or a maximum entropy distribution [7], [18], [21], [23]. While seemingly related, representativeness is not guaranteed by unexpectedness: adding a pattern  $X$  to our background knowledge does not ensure that, relative to  $X$ , the now-most-surprising pattern will be representative with regard to either pattern  $X$ , or to the whole population.

Another seemingly obvious relation that turns out to be much more subtle is that to fairness in classification. A truly representative pattern implies statistical parity with regard to the control variable, although it is worth noting that both Dwork and Mullainathan explicitly mention that statistical parity should not be equated with fairness, as it can potentially be “blatantly unfair” on an individual level [16], [26].

In recent work, Feldman et al. [9] studied the notion of “disparate impact”—a legal term that says that the probability ratio of treatment (e.g., job offer) for the different groups must be at least 0.8—and proposed as a general technique to remove disparate impact via data pre-processing. In other words, unlike our approach, the global distribution is changed to de-correlate sensitive and target attributes. Related as it may be, their work clearly fails to extend to local pattern mining, as in the latter, it does not suffice to model the global distribution.

Perhaps closest to our approach is the line of work by Calders et al. [6], who studied the goal of achieving statistical parity in classification with different methods, including naive bayes [5] and decision trees [14]. Kamishima et al. [15] consider a form of fairness that is related to statistical parity, although implicitly: during logistic regression a regularisation term is used, measuring the KL divergence between sensitive attribute and prediction. Although related, it is unclear whether

---

Dataset	Target $y$	Control $c$	$\alpha$	$ V $	$ P $
baseball	Salary	Fr.Ag.Elig.	1.0	16	268
gold	Evdw-Evdw0	Odd	1.0	19	12200
homicide	Victims	PerpRace	1.0	10	47236
students	G3	failures	0.5	31	366
wine	quality	colour	1.0	12	3198
abalone	Rings	Height	1.0	8	4144
aileron	Sa	RollRate	1.0	5	7108
airfoil	NoiseLevel	Displacement	1.0	5	1480
automp	Mpg	Cylinders	1.0	8	380
bike	registered	atemp	1.0	13	730
california	Med.Value	Latitude	0.5	8	20502
compactiv	usr	freeswap	0.7	21	8192
concrete	Strength	Age	1.0	8	562
elevators	Goal	DiffRollRate	0.3	18	16020
forestfires	Area	Month	0.6	12	394
house	Price	P14p9	0.3	16	22784
mortgage	30YRate	Mat.Rate3Y	0.8	15	1044
mv	Y	X6	0.9	10	40768
pole	Output	Att0	0.3	26	14586
puma32h	thetadd6	theta5	0.7	32	8192
stock	Company10	Company4	1.0	9	950
treasury	X1Rate	CMat.Rate3Y	0.4	15	1044
wankara	AvgTmp	MaxTemp	0.6	9	318
wizmir	AvgTmp	MaxTemp	0.5	9	1458

---

Table I: Used datasets, for qualitative (top) and quantitative (bottom) analysis. Listed are the number of attributes  $|V|$  and number of rows  $|P|$ , as well as running configurations: target and control variables, and approximation factor  $\alpha$ . The latter is decreased by 0.1 every time BRIG exceeds a timeout of 6 hours, or terminates due to exceeding 256GB of memory.

these methods can be utilised in the highly demanding B&B search, typically able to optimise over exponentially-sized discrete spaces of arbitrary subgroup descriptor languages.

Closest in terms of pattern mining, but relatively distant with regard to statistical parity, is the work by Pedreschi et al. [22] on discrimination-aware pattern mining. Instead of subgroups, the authors focus on mining association rules that may only include a sensitive item if this does not improve the confidence of the rule by more than  $\alpha$ .

## V. EXPERIMENTS

In this section we evaluate our extended impact function  $f$ , as well as RAWR, our implementation of its tight optimistic estimator for use by the B&B algorithm. We provide qualitative and quantitative demonstrations of superior representativeness in the discovered subgroups, and we also report runtime measurements on a variety of datasets.

For these tasks we implemented<sup>1</sup> both RAWR and the non-tight, representativeness oblivious BRIG, which we use as a baseline. We then run the B&B algorithm, using each of them.

### A. Mining more representative results

We now assess qualitatively and quantitatively the representativeness of the discovered subgroups, for different values of the  $\gamma$  parameter. We first study 5 datasets (retrieved from

<sup>1</sup> Our source code is available within the realKD tool [bitbucket.org/realKD/](http://bitbucket.org/realKD/).



the UCI ML repository [20] and the Murder Accountability Project <http://www.murderdata.org/>) which contain intuitively interpretable controls (Table I, top). To rule out the effect of unbalanced classes, and for our algorithm to be applicable, we stratify the datasets over the control classes; we then perform subgroup discovery and measure the  $f_r$  and  $f_{ct}$  scores of the discovered subgroups, as we increase the  $\gamma$  parameter (Fig. 4).

Obviously, a value of  $\gamma = 0$  corresponds to the representativeness-oblivious impact function  $f_{ct}$  of Eq. (1). Depending on the dataset and choice of  $y$  and  $c$ , the discovered subgroups for this case may be representative, although this is not guaranteed. However, as the  $\gamma$  parameter increases, the added  $f_r$  factor comes in effect, yielding consistently more representative subgroups (Figure 4a). As expected, the  $f_{ct}$  score may drop, demonstrating that  $\gamma$  controls the trade-off between the two factors (Figure 4b). At the same time, it is guaranteed that no score can be increased without the decrease of the other, by choosing a subgroup other than the discovered.

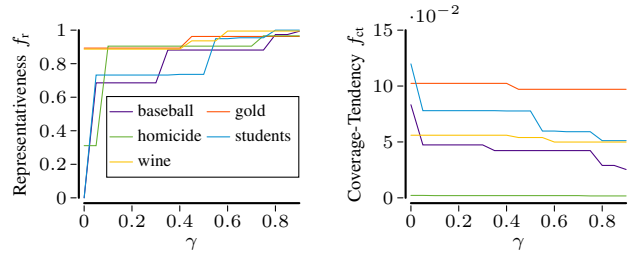
We next delve into the subgroups discovered in selected datasets. We first focus on the *Homicide* dataset, which tracks homicide cases, matched with background data on perpetrators and their victims, alongside the number of victims per case. We use the latter as a target variable to measure violence and seek to gain insight on attributes leading to increased violence, as captured by binary control variables. For each studied variable, we stratify the dataset and report the discovered subgroups (Table III), for increasing  $\gamma$  parameter.

We first consider the effect of the `Perpetrator Sex`. The subgroup discovered without the  $f_r$  term rediscovers the unsurprising fact that males are more violent than females. To uncover further potentially underlying factors, we use the `Perpetrator Sex` as a control variable and perform subgroup discovery using the controlled impact function. As  $\gamma$  parameter increases, the discovered subgroups hold for both male and female perpetrators, leading to the discovery that **Caucasian victims attract more violence**, and further that **no sex is more violent when it comes to older victims**.

In another example, we study an application for fair subgroup discovery. Consider that a baseball team decides to increase its players salary and seeks to find the factors that lead to higher income drawing experience from other team managements. At the same time, the raise should not be unfavourable to players who are contractually bound to one particular team, in contrast to the Free Agent eligible players, which might earn more lest they leave the team. Using the `FreeAgencyEligibility` variable as control, more objective criteria describing high salaries are discovered.

### B. Evaluating the performance of the proposed tight estimator

We now evaluate the performance of the RAWR implementation. To sample a broad variety of datasets, we used all of the regression datasets from the KEEL database [1] with a number of variables  $8 \leq |V| \leq 40$  (Table I, bottom). As a target variable we used the designated regression variable. To emulate a purported scenario of controlling for the main data trend, we use as control the first variable that appears in the



(a) Representativeness score  $f_r$ . (b) Coverage-tendency score  $f_{ct}$ .

Figure 4: Scores of subgroups discovered in the qualitative datasets (Table I, top). Tuning  $\gamma$  effectively controls the trade-off between Representativeness and coverage-tendency.

Dataset	$\gamma = 0.4$		$\gamma = 0.5$		$\gamma = 0.6$	
	RAWR	BRIG	RAWR	BRIG	RAWR	BRIG
gold	172	<b>101</b>	210	<b>99</b>	224	<b>121</b>
wine	296	<b>267</b>	349	<b>305</b>	375	<b>360</b>
house	14	<b>5</b>	13	<b>5</b>	17	<b>4</b>
stock	15	<b>8</b>	16	<b>10</b>	17	<b>12</b>
california	3	<b>2</b>	4	<b>2</b>	4	<b>2</b>
pole	4	<b>2</b>	3	<b>2</b>	3	<b>1</b>
airfoil	<b>9</b>	9	7	8	9	<b>9</b>
concrete	<b>4</b>	6	<b>5</b>	5	<b>6</b>	6
elevators	3	<b>3</b>	3	<b>3</b>	<b>2</b>	2
puma32h	<b>78</b>	84	<b>82</b>	83	<b>83</b>	85
baseball	<b>20</b>	22	<b>17</b>	21	<b>16</b>	19
bike	<b>5</b>	10	<b>5</b>	12	<b>6</b>	11
forestfires	<b>184</b>	272	<b>178</b>	186	<b>193</b>	217
homicide	<b>154</b>	219	<b>171</b>	247	<b>169</b>	236
aileron	<b>206</b>	317	<b>297</b>	486	<b>703</b>	797
compactiv	504	<b>450</b>	442	<b>349</b>	<b>1299</b>	3397
mv	<b>3877</b>	6837	<b>3243</b>	5273	<b>2300</b>	5026
students	<b>19</b>	175	<b>63</b>	2638	<b>126</b>	4768
autompg	<b>6</b>	3229	<b>6</b>	5577	<b>11</b>	10591
abalone	<b>869</b>	1307	<b>1883</b>	3876	<b>3639</b>	17575
wankara	<b>1</b>	116	<b>1</b>	2543	<b>1</b>	13977
mortgage	<b>56</b>	$\infty$	<b>198</b>	$\infty$	<b>12568</b>	$\infty$
treasury	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>	$\infty$
wizmir	5	<b>3</b>	2	1648	<b>3</b>	$\infty$

Table II: [Bold is better] Runtime comparison of RAWR and BRIG over different  $\gamma$  parameters for all datasets, sorted in increasing time difference. Using BRIG on the last 3 datasets exceeds our 256GB of memory, so results are not available.

subgroup descriptor discovered for  $\gamma = 0$ ; if this variable is real we discretise it around the median. Next, we stratify the dataset on the control variable. We start with an approximation factor of  $\alpha = 1$ , corresponding to exact computation; when all BRIG invocations for a dataset fail, due to either a runtime of more than 6 hours or exceeding 256GB of available memory, we decrease  $\alpha$  by 0.1 and repeat.

We assess the performance of our algorithm w.r.t. the number of searched nodes during each B&B invocation and also the needed runtime (Fig. 5); we set  $\gamma = 0.6$ , corresponding to a reasonably practical scenario. As our proposed optimistic estimator is tighter, it is yielding a significantly better pruning performance. What is more, our implementation seems to make use of the better pruning achieved, in order to attain running times that are comparable to those of BRIG, or in some

	$\gamma$	Subgroup describing $Q$	$f_r(Q)$	$f_{ct}(Q)$
<i>homicide</i>	Control: Perpetrator Sex			
	( 0, 0.09]	Crime=Murder, Vict.=White, Perp.=♂	0.00	0.002
	(0.09, 0.75]	Vict.=White	0.89	0.002
	[ 0.75, ...)	Vict.Age= ¬V.Lo, Vict.=White	0.99	0.001
<i>baseball</i>	Control: Perpetrator Race			
	( 0, 0.6]	Crime=Murder, Vict.=♀, Perp.= ¬V.Old	0.90	0.002
	[ 0.6, ...)	Crime=Murder, Vict.=♀, Perp.= ¬Old	0.98	0.002
	Control: Free Agent Eligibility			
( 0, 0.09]	OnBase= ¬V.Lo, F.A.Eligible= ✓	0.00	0.083	
(0.09, 0.33]	OnBase=HI	0.69	0.047	
(0.33, 0.8]	Batting= ¬V.Lo, OnBase= ¬Lo	0.88	0.042	
[ 0.8, ...)	Batting=¬V.Lo, OnBase= ¬Lo, Fr.Ag.= ✗	0.97	0.029	

Table III: Discovered subgroups for a varying  $\gamma$  parameter, for the datasets *homicide* (above) and *baseball* (below). Increasing  $\gamma$  produces more representative subgroups.

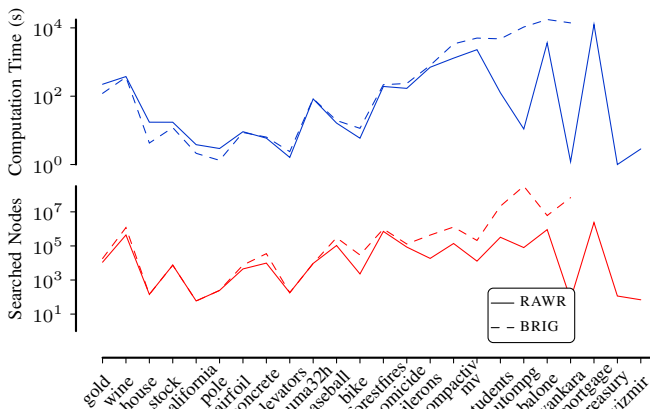


Figure 5: [Lower is better] Performance comparison of RAWR (solid) and BRIG (dashed), for runtime (blue) and searched nodes (red), with  $\gamma = 0.6$ . The datasets (x axis) are sorted in increasing time difference. Using BRIG on the last 3 datasets exceeds our 256GB of memory, so results are not available.

cases up to 4 orders of magnitude better. Further numerical results are reported in Table II, for a set of sensible weights  $\gamma \in \{0.4, 0.5, 0.6\}$ . These show a superiority of our estimator especially for higher  $\gamma$  values, where BRIG is less tight.

Furthermore, note that the number of nodes is a key factor contributing to the memory requirement of the B&B algorithm. As such, even for dataset on which the computation time of these implementations might be comparable, it is sometimes the case that the decreased number of nodes is enabling the computation using RAWR, where otherwise BRIG would exceed available memory, e.g., in the last 3 datasets of Table II.

## VI. DISCUSSION

Our introduced method guarantees the optimality of the results given the specified parameter, while optionally enabling a faster computation by relaxing the optimality guarantee.

The sole parameter  $\gamma$  of our method remains intuitive in its interpretation and possibly in its selection, regardless of

the input, with the zero value corresponding to a vanishing effect of our extension and a high value an increased weight of it. Nonetheless, not every dataset is equally sensitive to the intermediate values and the researcher is still required to make educated guesses based equally well on expert knowledge or resort to a trial and error scheme.

A further point of interest is the rate of convergence. Inheriting the weak points of the B&B algorithm, the worst case complexity is no better than exponential, although in no real world cases has this been observed. Additionally, there seems to be no good estimate on the difficulty of the dataset.

As a downside, our implementation of the tight estimator for our objective function requires a balanced, binary dataset. However, the case of binary classes is amongst the most usual, and the former can be solved by stratification as a workaround.

In the future we are planing to further broaden the application settings of our estimator to more than binary classes and non-balanced datasets, as well as to investigate the underlying causes of this inherently difficult optimisation problem.

## VII. CONCLUSION

We introduce the problem of representativeness aware subgroup discovery, where we want to discover subgroups that are exceptional with regard to the target variable, yet at the same time be to statistical parity with regard to the control variable. We show how we can achieve this by extending the typically used impact function by incorporating a tuneable representativeness factor. We propose a tight optimistic estimator for the newly representativeness aware impact function, and give an efficient algorithm to compute it in  $O(n \log n)$  time. Experiments show it may lead to orders of magnitude fewer node expansions, compared to the representative ignorant estimator, which is leveraged in similar orders of speedup. In future work we aim to extend our theory to nominal control variables, and studying exceptional representative model mining [8].

## REFERENCES

- [1] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, “KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework.” *Mult.-Val. Logic & Soft Comp.*, 2011.
- [2] M. Atzmueller, “Subgroup discovery,” *WIREs DMKD*, pp. 35–49, 2015.
- [3] M. Boley, B. R. Goldsmith, L. M. Ghiringhelli, and J. Vreeken, “Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery,” *DAMI*, pp. 1391–1418, 2017.
- [4] M. Boley and H. Grosskreutz, “Non-redundant subgroup discovery using a closure system,” in *ECML PKDD*. Springer, 2009, pp. 179–194.
- [5] T. Calders and S. Verwer, “Three naive Bayes approaches for discrimination-free classification,” *DAMI*, pp. 277–292, 2010.
- [6] T. Calders and I. Žliobaitė, “Why unbiased computational processes can lead to discriminative decision procedures,” in *Discrimination and Privacy in the Information Society*. Springer, 2013, pp. 43–57.
- [7] T. De Bie, “Maximum entropy models and subjective interestingness: an application to tiles in binary databases,” *DAMI*, pp. 407–446, 2011.
- [8] W. Duivesteijn, A. J. Feelders, and A. Knobbe, “Exceptional model mining,” *DAMI*, vol. 30, no. 1, pp. 47–98, 2016.
- [9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact,” in *KDD*. ACM, 2015, pp. 259–268.
- [10] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas, “Assessing data mining results via swap randomization,” *ACM TKDD*, pp. 167–176, 2007.



- [11] B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, and L. M. Ghiringhelli, "Uncovering structure-property relationships of materials by subgroup discovery," *New Journal of Physics*, vol. 19, no. 1, 2017.
- [12] H. Grosskreutz, S. Rüping, and S. Wrobel, "Tight optimistic estimates for fast subgroup discovery," in *ECML PKDD*, 2008, pp. 440–456.
- [13] S. Hanhijärvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila, "Tell me something I don't know: randomization strategies for iterative data mining," in *KDD*. ACM, 2009, pp. 379–388.
- [14] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning." IEEE, 2010, pp. 869–874.
- [15] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," in *ECML PKDD*, Sep. 2012, pp. 35–50.
- [16] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv:1609.05807*, 2016.
- [17] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Adv. in Knowl. Disc. & Data Min.*, 1996, pp. 249–271.
- [18] K.-N. Kontonasis and T. De Bie, "An information-theoretic approach to finding noisy tiles in binary databases," in *SDM*, 2010, pp. 153–164.
- [19] F. Lemmerich, M. Atzmueller, and F. Puppe, "Fast exhaustive subgroup discovery with numerical target concepts," *DAMI*, pp. 711–762, 2016.
- [20] M. Lichman, "UCI machine learning repository," <http://archive.ics.uci.edu/ml>, 2013.
- [21] M. Mampaey, J. Vreeken, and N. Tatti, "Summarizing data succinctly with the most informative itemsets," *ACM TKDD*, vol. 6, pp. 1–44, 2012.
- [22] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *KDD*. ACM, 2008, pp. 560–568.
- [23] N. Tatti, "Maximum entropy based significance of itemsets," *Knowl. Inf. Sys.*, pp. 57–77, 2008.
- [24] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in *PKDD*. Springer, 1997, pp. 78–87.
- [25] U. Yüceer, "Discrete convexity: convexity for functions defined on discrete spaces," *Discr. Appl. Math.*, pp. 297–304, 2002.
- [26] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations." *ICML13*, pp. 325–333, 2013.

## APPENDIX

In this section we provide proofs for our theoretical claims.

### A. Proof of Lemma 2: optimal c-t properties

Let  $Q_\mu^* \subseteq Q$  be the set attaining the highest  $f_{ct}$  value among those with cardinality  $\mu$ . We now reinterpret Eq. (5) as follows: the element  $\pi^{(\mu)}$  is equal to the index of the equi-count refinement set  $\mathcal{R}_{\pi^*}$  containing  $Q_\mu^*$ . Within all sets with a fixed cardinality  $f_c$  remains constant, and  $Q_\mu^*$  is the set with the maximal central tendency  $f_t$ ; we can then show, similar to Proposition 1, that the maximiser of  $f_t$  contains the topmost  $\mu$  target values. Altogether,  $\pi^{(\mu)}$  is exactly the class count of

$$Q_\mu^* := \arg \max_{|R|=\mu, R \subseteq Q} f_{ct}(R) = \arg \max_{|R|=\mu, R \subseteq Q} f_t(R) = \bigcup_{i=1}^{\mu} \{e_i\},$$

whose control class count is equal to the quantity in Eq (6).

To show (ii) we proceed as follows. Since  $Q_\mu^*$  contains the top- $\mu$  elements, we rewrite Eq. (4) as

$$(f_{ct} \circ \pi)(\mu) = f_{ct}(\pi^{(\mu)}) = \alpha_t \sum_{i=1}^{\mu} y_i - \alpha_c \mu,$$

with discrete derivatives

$$\begin{aligned} \Delta_\mu(f_{ct} \circ \pi) &= f_{ct}(\pi^{(\mu+1)}) - f_{ct}(\pi^{(\mu)}) = \alpha_t y_{\mu+1} - \alpha_c \\ \Delta_\mu^2(f_{ct} \circ \pi) &= \alpha_t (y_{\mu+1} - y_\mu) \leq 0, \end{aligned}$$

where the last inequality holds because  $y_\mu$  are decreasing. The negativity of the second discrete derivative, shows the concavity of the sequence.

### B. Proof of Proposition 3: sufficient search triangle

This proof involves partitioning the CCS in compact regions surrounding the SST, within which the monotonicity of the factors  $f_{ct}^Q$  and  $f_r^Q$  remains constant, when computed along any horizontal or vertical sequences. All the sequences formed in this way increase toward the region boundary intersecting with the SST, and so no maximiser of  $f^Q$  can lie within these regions, except on the intersection of the region and the SST.

To show the above, we study both factors, starting with  $f_{ct}^Q$ .

**Lemma 5** (Domination of the  $f_{ct}^Q$  factor). *The impact value computed along any horizontal sequence is increasing until the abscissa  $\pi_1^*$  of the optimal c-t point, and decreasing afterwards. Similarly, the impact value computed along any vertical sequence is increasing until the ordinal  $\pi_2^*$  of the optimal c-t point, and decreasing afterwards. Formally,*

$$\begin{aligned} f_{ct}^Q(\mathbf{I} + e_k) &\geq f_{ct}^Q(\mathbf{I}), & I_k < \pi_k^* \\ f_{ct}^Q(\mathbf{I} + e_k) &\leq f_{ct}^Q(\mathbf{I}), & I_k \geq \pi_k^* \end{aligned}$$

*Proof.* Denote the **optimal c-t path index**  $\mu^*$  to be the index within the c-t path sequence attaining the maximum c-t value,  $\pi^{(\mu^*)} = \pi^*$ , so that

$$\begin{aligned} f_{ct}(\pi^{(\mu+1)}) &\geq f_{ct}(\pi^{(\mu)}) & \mu < \mu^* \\ f_{ct}(\pi^{(\mu+1)}) &\leq f_{ct}(\pi^{(\mu)}) & \mu \geq \mu^* \end{aligned}, \quad (7)$$

due to the concavity of the sequence  $(f^Q \circ \pi)(\mu)$ .

However, for any two consecutive points on the path, we can compute  $f_{ct}(\pi^{(\mu+1)}) - f_{ct}(\pi^{(\mu)}) = \alpha_t y_{\mu+1} - \alpha_c$ , which combined with Eq. (7) yields

$$\begin{aligned} \alpha_t y_{\mu+1} - \alpha_c &\geq 0 & \mu < \mu^* \\ \alpha_t y_{\mu+1} - \alpha_c &\leq 0 & \mu \geq \mu^* \end{aligned}. \quad (8)$$

Moreover, using Eq. (4) we can express the  $f_{ct}^Q$  value of the point next to  $\mathbf{I}$  in CCS along dimension  $k$  as

$$f_{ct}^Q(\mathbf{I} + e_k) = \alpha_t \sum_{k=1}^2 \left( \sum_{i=1}^{I_k} y_i^{(k)} + y_{I_k+1}^{(k)} \right) - \alpha_c (I_1 + I_2),$$

and so the difference between the  $f_{ct}^Q$  values of these neighbouring points becomes

$$f_{ct}^Q(\mathbf{I} + e_k) - f_{ct}^Q(\mathbf{I}) = \alpha_t y_{I_k+1}^{(k)} - \alpha_c, \quad (9)$$

which is the quantity whose sign we study. According to Eq. (6), however,  $\pi$  is a sequence of single step increases  $\pi^{(\mu+1)} - \pi^{(\mu)} = e_{c_\mu}$ , starting from the empty count  $\mathbf{0}$ . In other words, the  $\mu$ -th element  $\pi^{(\mu)}$  of the sequence increases this class count that matches the class of the item in  $Q$  with the next greatest target value. This implies that at the optimal c-t path index  $\mu^*$ , the optimal c-t path count  $\pi^* = \pi^{(\mu^*)}$  per class amounts exactly to the number of items with the same control class and greater target value. Moreover, for each  $I_k \geq \pi_k^*$  there exists a  $\mu \geq \mu^*$  such that  $y_\mu = y_{I_k+1}^{(k)}$ , and similarly for each  $I_k \leq \pi_k^*$  there exists a  $\mu \leq \mu^*$  such that  $y_\mu = y_{I_k+1}^{(k)}$ . We can now combine the two equations Eq. (9), and Eq. (8), to show the claim of the lemma.  $\square$

We now proceed to show a similar behaviour of the  $f_r$  factor.

**Lemma 6** (Total Variation domination). *The composition of  $f_r$  with every horizontal sequence  $\mathbf{h}_i$ ,  $i = 0, \dots, n_1(Q)$ , and every vertical sequence  $\mathbf{v}_i$ ,  $i_1 = 0, \dots, n_1(Q)$  forms the sequences  $(f_r \circ \mathbf{h}_i)(\tau)$  and  $(f_r \circ \mathbf{v}_i)(\tau)$ ; these are (i) unimodal, (ii) attain a maximum at their intersection  $(i, i)^T$  with the equi-representativeness ray  $a(1, 1)^T$ ,  $a \geq 0$ , and (iii) they are concave for the indices  $\tau = 0, \dots, i$ .*

*Proof.* We first focus on the horizontal sequences  $(f_r \circ \mathbf{h}_i)(\tau)$  for  $i = 0, \dots, n_2(Q)$  and  $\tau = 0, \dots, m$ . Notice that the  $d_{TV}(\mathbf{I})$  vanishes on the equi-representativeness ray  $a(1, 1)^T$ , that is, when  $I_1 = I_2$ . Since the horizontal sequence  $\mathbf{h}_i$  has a fixed ordinal of  $i$ , the previous condition yields  $\tau = i$ , which shows the correctness of (ii).

To prove the rest of this lemma, we study the continuous analogue of  $(f_r \circ \mathbf{h}_i)(\tau)$

$$\tilde{f}_r(t) := 1 - \left| \frac{1}{2} - \frac{t}{t+i} \right|, \quad t > 0,$$

which has first and second derivatives

$$\begin{aligned} \tilde{f}_r'(t) &= \text{sign} \left( \frac{1}{2} - \frac{t}{t+i} \right) \frac{i}{(t+i)^2} \\ \tilde{f}_r''(t) &= -2 \text{sign} \left( \frac{1}{2} - \frac{t}{t+i} \right) \frac{i}{(t+i)^3} \end{aligned} \quad t \neq i.$$

The sign of both quantities is controlled by the sign factor, which is negative when  $t < i$  and positive when  $t > i$ , and so we can reach the conclusion that  $\tilde{f}_r'$  is increasing concave for  $t < i$  and decreasing convex for  $t > i$ . Since  $(f_r \circ \mathbf{v}_i)(\tau) = \tilde{f}_r(t)$ , the above properties transfer to the discrete sequence  $(f_r \circ \mathbf{v}_i)$ , as well. For vertical sequences, the symmetric argumentation can be used.  $\square$

Combining the two domination lemmas 5 and 6, we can now prove the sufficiency of SST, by showing that every point outside the SST is dominated by one within  $\mathcal{T}(Q)$ . For this we distinguish two cases, depending on whether the c-t optimal point is above or below the maximum representativeness line.

Assume the optimal c-t point lies above the maximum  $f_r^Q$  ray. The point  $\mu^*$ , along with the maximum representativeness ray, partition the CCS in the 6 regions shown in Fig 6, each of which has a non empty intersection with  $\mathcal{T}(Q)$ . We now show that the maxima of  $f$  over all the points in these regions, lie on this intersection, and therefore also in the SST.

We first study  $A_{SW}$ : the points on the diagonal maximise  $f_r^Q$ , while at the same time  $f_{ct}^Q$  is dominated by the SST point  $(\pi_1^*, \pi_1^*)$ , therefore maximising  $f$  altogether. Similarly within  $A_{NE}$ , we can show that  $f$  is maximised by  $(\pi_2^*, \pi_2^*) \in \mathcal{T}(Q)$ .

Within regions  $A_W$  and  $A_N$ , both terms increase along each west-to-east and north-to-south path, respectively; these paths lead to a point of  $\mathcal{T}(Q)$  that dominates all the rest on the traversed path. Within  $A_{SW}$ , each west-to-east path ends up in a point of  $A_N$ , which is itself dominated by a point of SST.

Finally, every south-to-north path of  $A_{SE}$  leads to either a point of  $\mathcal{T}(Q)$  directly, or to one in the dominated  $A_{NE}$ . We thus showed that no point of  $\mathcal{I}(Q) \setminus \mathcal{T}(Q)$  can maximise  $f$ .

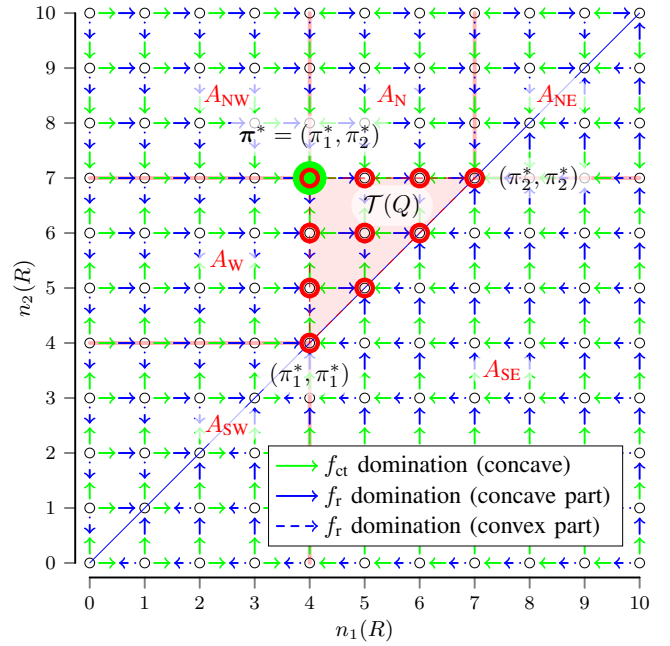


Figure 6: Domination relations for a c-t optimal point  $\pi^*$  above the maximum  $f_r^Q$  ray: the arrows point to the greater factor value. The  $\mathcal{T}(Q)$  partitions  $\mathcal{I}$  in the 6 marked areas.

We work likewise if  $\pi^*$  lies below the maximum  $f_r^Q$  ray.

#### C. Proof of Proposition 4: concavity of $f^Q$ along sequence

To prove this statement we employ the concavity of the sequences formed as the two factors  $f_{ct}^Q$  and  $f_r^Q$  are computed along the horizontal and vertical sequences. We first treat the horizontal sequences, along which the entire  $f_{ct}^Q \circ \mathbf{h}_{i_2}$  is concave, and so is  $(f_r^Q \circ \mathbf{h}_{i_2})(\mu)$  for the indices  $\mu = 0, \dots, i_2$ , according to Lemmata 2 and 5, respectively.

Additionally, all factors are positive (or can be made so by adding an appropriate constant term) and so, raising the elements of the sequences to a power in  $(0, 1]$  preserves concavity. Multiplying the two re-weighted sequences yields

$$((f_{ct}^Q \circ \mathbf{h}_{i_2})^\gamma (f_r^Q \circ \mathbf{h}_{i_2})^{1-\gamma})(\mu) = (f^Q \circ \mathbf{h}_{i_2})(\mu),$$

which is concave as the multiplication of two concave, positive sequences, therefore showing the concavity of the sequence of impact function values computed along the specified horizontal sub-sequence. Similarly we can work for vertical sequences.

Note that in our analysis we seamlessly interchange continuous and discrete convexity definitions. This is enabled by the uni-variate nature of the functions involved, since their discrete counterparts corresponds to sampling on regular intervals. Indeed, on one hand it can be shown that regular sampling of a uni-variate convex function yields a convex sequence [25]. As a sufficiently applicable inverse for our needs, we can show that for any convex sequence, there exists at least one convex function with the same values at the sampled points and continuous second derivative; one such function results from cubic spline interpolation fitted on the sequence values.