Optimal Subgroup Discovery in Purely Numerical Data

Alexandre Millot, Rémy Cazabet, Jean-François Boulicaut

Optimal Subgroup Discovery in Purely Numerical Data

Alexandre Millot[1], Rémy Cazabet[2], and Jean-François Boulicaut[1]

[1] Univ de Lyon, CNRS, INSA Lyon, LIRIS, UMR5205, F-69621, France
{firstname.lastname}@insa-lyon.fr
[2] Univ de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
{firstname.lastname}@univ-lyon1.fr

Abstract. Subgroup discovery in labeled data is the task of discovering patterns in the description space of objects to find subsets of objects whose labels show an interesting distribution, for example the disproportionate representation of a label value. Discovering interesting subgroups in purely numerical data - attributes and target label - has received little attention so far. Existing methods make use of discretization methods that lead to a loss of information and suboptimal results. This is the case for the reference algorithm SD-Map*. We consider here the discovery of optimal subgroups according to an interestingness measure in purely numerical data. We leverage the concept of closed interval patterns and advanced enumeration and pruning techniques. The performances of our algorithm are studied empirically and its added-value w.r.t. SD-Map* is illustrated.

Keywords: Pattern Mining · Subgroup Discovery · Numerical Data

1 Introduction

Mining purely numerical data is quite popular. It concerns data made of objects described by numerical attributes, and one of these attributes can be considered as a target label. We can then choose to learn models to predict the value of the label for new objects, or we can apply subgroup discovery methods [14,22] that is the focus of this paper. Subgroup discovery aims at discovering subsets of objects - known as subgroups - described by interesting descriptions according to a quality measure calculated on the target label. A quality measure has to capture discrepancies in the target label distribution between the selected subset of objects and the overall dataset. A large panel of exhaustive [1,10] and heuristic [16,5] subgroup discovery algorithms have been proposed so far. Most of these approaches consider a set of nominal attributes with a binary label. Regarding numerical attributes, a few approaches [11,19] that avoid the use of basic discretization techniques have been introduced. However, to the best of our knowledge, we lack from a method that would support an exhaustive search and thus the possibility to guarantee the computation of a global optimum for the quality measure without the use of discretization in some form or other. When

considering numerical target labels, [15] introduced relevant quality measures as well as the SD-Map$\star$ reference algorithm. Notice however that SD-Map$\star$ requires the prior discretization of the numerical attributes.

The guarantee to discover an optimal subgroup in purely numerical data is a useful task and we now motivate it for our ongoing research project. We are currently working on optimization methods for urban farms (e.g., AeroFarms, Infarm, FUL[1]). In that setting, plant growth recipes involve many numerical attributes (temperature, hydrometry, CO_2 concentration, etc) and a numerical target label (the yield, the energy consumption, etc). Our goal is to mine the recipe execution records (i.e., the collected measures) to discover the characteristics of an optimized growth. In expert hands, such characteristics can be exploited to define better recipes. In such a context, the guaranteed discovery of the optimal subset of recipes with respect to the target label is more relevant than the heuristic discovery of the k best subgroups with no optimality guarantee. Preliminary results on simulated crops are given in this paper.

To achieve the search for optimality, we decided to search the space of interval patterns as defined in [13]. Our main contribution consists in an algorithm that exhaustively enumerates all the interval patterns. Our approach (i) exploits the concept of closure on the positives adapted to a numerical setting to operate in a subspace (ii) uses a new faster tight optimistic estimate that can be applied for several quality measures (iii) uses advanced pruning techniques (forward checking, branch reordering). The result is the efficient algorithm OSMIND for an optimal subgroup discovery in purely numerical data without prior discretization of the attributes. Section 2 formalizes our mining task. In Section 3, we discuss related work. We detail our contributions in Section 4 before an empirical evaluation in Section 5. Section 6 briefly concludes.

2 Problem Definition

Purely numerical dataset. A purely numerical dataset (G, M, T) is given by a set of objects G, a set of numerical attributes M and a numerical target label T. In a given dataset, the domain of any attribute $m \in M$ is a finite ordered set denoted D_m. In this context, $m(g) = d$ means that d is the value of attribute m for object g. The domain of label T is also a finite ordered set denoted D_T. $T(g) = v$ means that v is the value of label T for object g. Fig. 1 (left) is a purely numerical dataset made of two attributes ($M = \{m_1, m_2\}$) and a target label T. A subgroup p is defined by a pattern, i.e., its intent or description, and the set of objects from the dataset where it appears, i.e., its extent, denoted $ext(p)$.

Interval patterns, extent and closure. Given a purely numerical dataset (G, M, T), an interval pattern p is a vector of intervals $p = \langle [b_i, c_i] \rangle_{i \in \{1,\ldots,|M|\}}$ where $b_i, c_i \in D_{mi}$, and each interval is a restriction on an attribute of M, and $|M|$ is the number of attributes. An object $g \in G$ is in the extent of an interval

[1] https://aerofarms.com/, https://infarm.com/, http://www.fermeful.com/.

	m_1	m_2	T
g_1	1	1	15
g_2	1	2	30
g_3	2	2	60
g_4	3	2	40
g_5	3	3	70
g_6	4	3	85

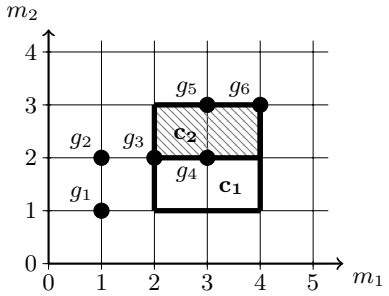


Fig. 1. (**left**) A purely numerical dataset. (**right**) Non-closed ($c_1 = \langle [2,4],[1,3] \rangle$, non-hatched) and closed ($c_2 = \langle [2,4],[2,3] \rangle$, hatched) interval patterns.

pattern $p = \langle [b_i, c_i] \rangle_{i \in \{1,...,|M|\}}$ iff $\forall i \in \{1,...,|M|\}, m_i(g) \in [b_i, c_i]$. Let p_1 and p_2 be two interval patterns. $p_1 \subseteq p_2$ means that p_2 encloses p_1, i.e., the hyper-rectangle of p_1 is included in that of p_2. It is said that p_1 is a specialization of p_2. Given an interval pattern p and its extent $ext(p)$, p is defined as *closed* if and only if it represents the most restrictive pattern (i.e., the smallest hyper-rectangle) that contains $ext(p)$. For example, in the data from Fig. 1 (left), the domain of m_1 is $\{1,2,3,4\}$ and $\langle [2,4],[1,3] \rangle$ is the interval pattern that denotes a subgroup whose extent is $\{g_3, g_4, g_5, g_6\}$. Fig. 1 (right) depicts the same dataset in a cartesian plane as well as a comparison between a non-closed (c_1) and a closed (c_2) interval pattern.

Quality measure and optimistic estimate. Considering a purely numerical dataset (G, M, T), the interestingness of each interval pattern is measured by a numerical value. Usually, the value quantifies the discrepancy in the target label distribution between the overall dataset and the extent of the considered interval pattern. We consider here the family of quality measures based on the mean introduced in [15]. Given an interval pattern p, its quality is given by:

$$q^a_{mean}(p) = |ext(p)|^a \times (\mu_{ext(p)} - \mu_{ext(\emptyset)}), a \in [0,1]$$

with $\mu_{ext(p)}$ the mean of the target label for p, $\mu_{ext(\emptyset)}$ the mean of the target label for the overall dataset, $|ext(p)|$ the cardinality of $ext(p)$ and a a parameter that controls the number of objects in the subgroups.

Given an interval pattern p and a quality measure q, an optimistic estimate for q, denoted as bs_q, is a function that gives an upper bound for the quality of all specializations of p. Formally, $\forall s \subseteq p : q(s) \leq bs_q(p)$. Optimistic estimates are used to prune the search space: if an interval pattern optimistic estimate is lower than the required minimal quality, it is useless to consider its specializations.

Optimal subgroup. Let (G, M, T) be a purely numerical dataset, q a quality measure and P the set of all interval patterns of (G, M, T). An interval pattern is said to be optimal iff $\forall p' \in P : q(p') \leq q(p)$. Notice that several subgroups

can have the same optimal quality. In this paper, we return the first one found by the algorithm.

3 Related Work

Although we are not aware of previous proposals for an optimal subgroup discovery in purely numerical data, related topics have been seriously investigated. Traditionally, subgroup mining has been mainly concerned with nominal attributes and binary target labels. To deal with numerical data, prior discretization of the attributes [6,8] is then required. Numerical target labels can also be discretized [18]. However, discretization generally involves a loss of information such that we cannot guarantee the optimality of the returned subgroups. [2] introduced the concept of *Quantitative Association Rules* where a rule consequent is the mean or the variance of a numerical attribute. A rule is then defined as interesting if its mean or variance significantly deviates from that of the overall dataset. Later on, [21] proposed an extension of such rules called the *Impact Rules*. These methods, however, cannot perform an exhaustive enumeration of subgroups and therefore provide no guarantee for an optimal subgroup discovery. A recurring issue with exhaustive pattern enumeration algorithms is the size of the search space which is exponential as a function of the number of attributes. Fortunately, the search space can be pruned thanks to optimistic estimates [22,12]. [15] introduces a large panel of quality measures and optimistic estimates for an exhaustive mining with numerical target labels. A few approaches have been proposed to tackle numerical attributes [11,16,19]. However, these methods always involve the use of discretization techniques. When dealing with exhaustive search in numerical data, we find the `MinIntChange` algorithm [13] based on constructs from Formal Concept Analysis [7]. It enables an exhaustive mining of frequent patterns - not subgroups - in numerical data without prior discretization. The use of closure operators and equivalence classes [20,9,4,10] is a popular solution to reduce the size of the subgroup search space. [3] introduced an anytime subgroup discovery algorithm in numerical data for binary target labels by revisiting the principles of `MinIntChange`. We also want to leverage closure operators, optimistic estimates and the enumeration strategy of `MinIntChange` for an optimal subgroup discovery in purely numerical data though our mining task is different from the task in [3].

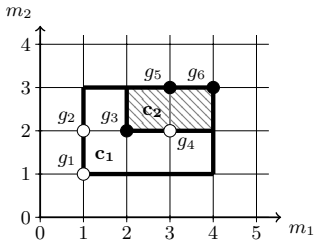
4 Optimal Subgroup Discovery

4.1 Closure On The Positives

The closure operator on interval patterns introduced in [13] has been extended to closure on the positives for binary labels in [3].

Definition 1. *Let $p \in P$ be an interval pattern, $p' \subseteq p$ a second interval pattern, and T a binary target label. An object is said to be positive if its label value is*

	m_1	m_2	T	T_b
g_1	1	1	15	−
g_2	1	2	30	−
g_3	2	2	60	+
g_4	3	2	40	−
g_5	3	3	70	+
g_6	4	3	85	+



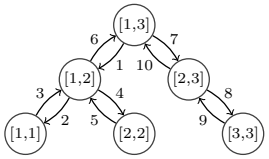


Fig. 2. (**left**) Purely numerical dataset with binary label (T_b). (**center**) Closed ($c_1 = \langle [1,4],[1,3] \rangle$, non hatched) and closed on the positives ($c_2 = \langle [2,4],[2,3] \rangle$, hatched) interval patterns. (**right**) Depth-first traversal of D_{m_2} using minimal changes.

that of the class we want to discriminate, and negative in the opposite case. Let $ext(p)^+$ be the subset of objects of $ext(p)$ whose label T is positive. p' is said to be closed on the positives if it is the most restrictive pattern enclosing $ext(p)^+$. If q is the quality measure, we have $q(p) \leq q(p')$.

For all subgroups $p \in P$, if all negative objects which are not in the extent of p' are removed from the extent of p, then the subgroup quality cannot decrease. Note that closed on the positives are a subset of closed patterns.

The concept of closed on the positives for binary target labels can be extended to numerical target labels for a set of quality measures, including q^a_{mean}. We transform the numerical label into a binary label: objects whose label value is strictly higher (resp. lower or equal) than the mean of the dataset are defined as positive (resp. negative). Note that the quality measure is computed on the raw numerical label. The binarisation is only used to improve search space pruning and it does not lead to a loss of information concerning the resulting patterns (i.e., the optimal subgroup discovery without discretization is guaranteed). Fig. 2 (left) is the dataset of Fig. 1 with label T (mean = 50) transformed into the binary label T_b. Fig. 2 (center) depicts the dataset of Fig. 2 (left) in a cartesian plane and a comparison between a closed (c_1) and a closed on the positives (c_2) interval pattern. We separate the case where the subgroup quality is positive from the case where it is negative. Given a subgroup of positive quality, we can prove that its quality is always higher or equal if all negative objects not in the closure on the positives are removed.

Theorem 1. *Let p be an interval pattern, q^a_{mean} a set of quality measures, p^+ the closure on the positives of p such that $p^+ \subseteq p$, and $q^a_{mean}(p) \geq 0$, then $q^a_{mean}(p^+) \geq q^a_{mean}(p),\ a \in [0,1]$.*

Proof. Let $ext(p)$ be the extent of p, $ext(p)^+$ the extent of p^+, $ext(p)^- = ext(p) \setminus ext(p)^+$ the set of negative objects of $ext(p)$ not in $ext(p)^+$, and $T(i)$ the target label value for Object i. For shorter notation, we define $e = ext(p)$ and $\theta = ext(\emptyset)$. We prove that:

$$|e^+|^a \times (\mu_{e^+} - \mu_\theta) \geq |e|^a \times (\mu_e - \mu_\theta) \tag{1}$$

Which can be transformed into:

$$|e^+|^a \times \frac{\sum_{i \in e^+}(T(i) - \mu_\theta)}{|e^+|} \geq |e|^a \times \frac{\sum_{i \in e}(T(i) - \mu_\theta)}{|e|} \tag{2}$$

$$|e^+|^a \times \frac{\sum_{i \in e^+}(T(i) - \mu_\theta)}{|e^+|} \geq (|e^+| + |e^-|)^a \times \frac{\sum_{i \in e^+}(T(i) - \mu_\theta) + \sum_{i \in e^-}(T(i) - \mu_\theta)}{|e^+| + |e^-|} \tag{3}$$

By construction, we know that $\sum_{i \in e^+}(T(i) - \mu_\theta) \geq 0 \geq \sum_{i \in e^-}(T(i) - \mu_\theta)$. The rest of the proof follows the same as [15]. We deduce that for any subgroup verifying $q^a_{mean}(p) \geq 0$, the closure on the positives always leads to a subgroup of equal or higher quality.

The case of a negative quality subgroup is more complex since the closure on the positives can lead to a decrease of the subgroup quality. We prove that objects which are not in the closure on the positives can never be part of the best subgroup specialization.

Theorem 2. *Let p be an interval pattern, p^+ the closure on the positives of p such that $p^+ \subseteq p$ and $ext(p)^+$ its extent with $|ext(p)^+| > 0$. Let $ext(p)^- = ext(p) \setminus ext(p)^+$ be the set of negative objects of $ext(p)$ not in $ext(p)^+$, and q^a_{mean} a set of quality measures with $q^a_{mean}(p) < 0$: No object in $ext(p)^-$ can be part of the best specialization of p.*

Proof. We hypothesize that there exists an object in $ext(p)^-$, denoted i^-, which belongs to the best specilization of p, denoted p_{top}. By construction, $q^a_{mean}(p_{top}) > 0$ (since $|ext(p)^+| > 0$). Let p^+_{top} be the closure on on the positives of p_{top}. By construction, we know that i^- isn't part of the extent of p^+_{top} (since i^- doesn't belong to p^+). Yet, according to Theorem 1, we have $q^a_{mean}(p^+_{top}) \geq q^a_{mean}(p_{top})$. We deduce that i^- doesn't belong to the best specialization of p.

4.2 Tight Optimistic Estimate

We now introduce a new tight optimistic estimate for the family of quality measures q^a_{mean}. An optimistic estimate is said to be tight, if, for any subgroup of the dataset, there is a subset of objects of the subgroup whose quality is equal to the value of the subgroup optimistic estimate. Note that the subset does not need to be a subgroup. It is possible to derive a tight optimistic estimate for the quality measures q^a_{mean} by considering each object of a subgroup only once.

Definition 2. *Let p be an interval pattern, and $S_i \subseteq ext(p)$ the subset of objects of $ext(p)$ containing the i objects with the highest label value. Then, as defined in [15], a tight optimistic estimate for q^a_{mean} is given by:*

$$bss^a_{mean}(p) = max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)|})), a \in [0, 1]$$

We can derive a better optimistic estimate by focusing on positive objects only.

Theorem 3. *Let p be an interval pattern and $ext(p)^+$ the set of objects from the extent of p whose label value is higher than the mean of the dataset. Let $S_i \subseteq ext(p)^+$ be the subset of objects containing the i objects with the highest label value. A new tight optimistic estimate for q^a_{mean} is given by:*

$$\overline{bss}^a_{mean}(p) = max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)^+|})), a \in [0, 1]$$

Proof. We need to prove that:

$$\overline{bss}^a_{mean}(p) \geq bss^a_{mean}(p), a \in [0, 1]$$

In other words, we need to show that: $\forall S_i \subseteq ext(p), q^a_{mean}(S_i^+) \geq q^a_{mean}(S_i)$ with S_i^+ the subset of positive objects of S_i. In [15], it is proven that no negative object belongs to the best subgroup's subset of objects for the quality measures q^a_{mean}. It follows logically that for any subset S_i, removing the negative objects can not lower its quality. Thus, we have

$$\forall S_i \subseteq ext(p), q^a_{mean}(S_i^+) \geq q^a_{mean}(S_i)$$

We deduce that:

$$max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)^+|})) \geq max(q^a_{mean}(S_1), ..., q^a_{mean}(S_{|ext(p)|})), a \in [0, 1]$$

Thus, $\overline{bss}^a_{mean}(p)$ is a tight optimistic estimate for q^a_{mean}.

4.3 Algorithm

We introduce `OSMIND`, a *depth first search* algorithm for an optimal subgroup discovery. It computes closed on the positives interval patterns coupled with the use of tight optimistic estimates and advanced search space pruning techniques. The pseudocode is available in Algorithm 1.

To guarantee an optimal subgroup discovery, we adopt the concept of minimal change from `MinIntChange` that ensures an exhaustive enumeration of all interval patterns (see Fig. 2 (right) for an example with one attribute). A right minimal change consists in replacing the right bound of an interval by the current value closest lower value in the domain of the corresponding attribute. Following the same logic, a left minimal change consists in replacing the left bound by the closest higher value. The search starts with the minimal interval pattern that covers all the objects of the dataset. The main idea in procedure `RECURSION` is to apply consecutive left or right minimal changes until obtaining an interval whose left and right bounds have the same value for each interval of the minimal interval pattern. If so, the algorithm backtracks until finding a pattern on which a minimal change can be applied. We leverage the concept of closure on the positives adapted to numerical labels to significantly reduce the number of candidate interval patterns. After each minimal change (Line 4), instead of evaluating the resulting interval pattern, we compute and evaluate the corresponding closed on the positives interval pattern (Line 5). When carrying out an exhaustive search of all closed on the positives interval patterns, a given

Algorithm 1 OSMIND algorithm

```
1: function OSMIND( )
2:     Initialize(minimal_interval_pattern, optimal_pattern)
3:     RECURSION(minimal_interval_pattern, 0)
4:     return optimal_pattern
5: end function
1: procedure RECURSION(pattern, attribute)
2:     for i = attribute to nb_attributes − 1 do
3:         for elem in {right, left} do
4:             pattern ← minimalChange(pattern, i, elem)
5:             closed_pattern ← computeClosureOnThePositives(pattern)
6:             if isCanonical(closed_pattern) then
7:                 if tightOptEst(closed_pattern) > quality(optimal_pattern) then
8:                 store(closed_pattern, i) end if
9:                 if quality(closed_pattern) > quality(optimal_pattern) then
10:                optimal_pattern ← closed_pattern end if
11:            end if
12:        end for
13:    end for
14:    for each element stored ordered by optimistic estimate value do
15:        if tightOptEst(element.pattern) > quality(optimal_pattern) then
16:        RECURSION(element.pattern, element.attribute) end if
17:    end for
18: end procedure
```

interval pattern can be generated multiple times. To avoid this redundancy and to ensure the unicity of the pattern generation, a popular solution is the use of a canonicity test. In the case of interval patterns, the canonicity test verifies that the closure operation did not lead to a change on an interval preceding the interval on which the minimal change has been applied (Line 6). However, the successive application of left or right minimal changes on an interval can also lead to multiple generations of the same interval pattern. A solution is to use a constraint on the minimal changes. After a right minimal change, a right or left minimal change can be applied. However, a left minimal change must always be followed by a left minimal change. We also exploit advanced pruning techniques to reduce the size of the search space. This can be done through the use of a tight optimistic estimate of the quality of a closed on the positives interval pattern specializations. For each subgroup, an optimistic estimate is derived (Line 7), and, if it is lower than the best subgroup quality, the search space is pruned by discarding every specialization of this interval pattern. Our second implemented technique is the coupling of *forward checking* and *branch reordering*. Given an interval pattern, the set of all its direct specializations (application of a right or left minimal change on each interval) are computed - forward checking - and those whose optimistic estimate is higher than the best subgroup are stored (Line 8). Branch reordering by descending order of the optimistic estimate value is then carried out (Line 14). Branch reordering enables to explore the most

promising parts of the search space first. It also enables a more efficient pruning by raising the minimal quality earlier.

5 Empirical Validation

We consider 7 purely numerical datasets described in Table 1. `SD-Map*` implementation is available within the VIKAMINE system[2]. The first 5 datasets (Bolt, Basketball, Airport, Body Temp and Pollution) originate from the Bilkent[3] repository. The other 2 datasets (RecipesA and RecipesB) are simulations of plant growth that we generated using the specialized environment Python Crop Simulation Environment PCSE[4]. Each growth simulation is described by a set of numerical attributes - the growth conditions (e.g., temperature, CO_2) - and a numerical target label - the yield at the end of the growth cycle. Here, a plant growth is split into several time periods of equal length called *stages*. Table 2 depicts simplified examples of plant growth simulations generated with PCSE.

Table 1. Datasets and their characteristics: number of attributes, number of objects and size of the search space.

Dataset	Attr	Obj	|P|
Bolt	8	40	8.7×10^9
Basketball	4	96	2.3×10^{11}
Airport	4	135	7.1×10^{15}
Body Temp	2	130	1.8×10^3
Pollution	15	60	1.7×10^{42}
RecipesA	9	100	5.1×10^{18}
RecipesB	9	1000	5.1×10^{18}

Table 2. Plant growth split in 2 stages (P1 and P2), 2 attributes (temperature and CO_2), and a target label (yield).

R	T^{P1}	CO_2^{P1}	T^{P2}	CO_2^{P2}	Y
r_1	18	800	24	1000	5
r_2	22	1000	27	950	6
r_3	27	1200	28	650	7
r_4	19	600	17	800	3
r_5	24	500	23	450	9
r_6	16	750	19	1300	2
r_7	30	1100	25	900	8

Performance improvements provided by our contributions are summarized in Table 3. Performances of the closure on the positives operator are compared to those of a simple closure operator (Section 2). For each dataset, we compare the number of evaluated subgroups before finding the optimal one for the quality measure q^a_{mean} with $a = 0.5$ and $a = 1$. In all the cases, the closure on the positives is significantly more efficient. In fact, our method enables to divide the number of considered subgroups by an average of more than 20. We now study the potential performance improvement - in terms of execution time in seconds - provided by our new tight optimistic estimate. We compare it to the tight optimistic estimate from [15] on all the datasets with the same quality measures. Our optimistic estimate is more efficient in all cases and it provides an execution time decrease of up to 30%.

[2] `http://www.vikamine.org/`
[3] `http://funapp.cs.bilkent.edu.tr/DataSets/`
[4] `https://pcse.readthedocs.io/en/stable/index.html`

Let us discuss the added-value of OSMIND w.r.t. SD-Map⋆, i.e., the reference algorithm for an exhaustive strategy with numerical target labels. We compare the quality of the best found subgroup with each method on the first 5 datasets of Table 1 when using the quality measure q^a_{mean} with $a = 0.5$. Regarding SD-Map⋆, a prior discretization of numerical attributes is needed. To obtain fair results, we evaluate several discretization techniques with different numbers of cut-points (2, 3, 5, 10, 15 and 20) for SD-Map⋆ and we retain only the best solution that is compared to the OSMIND results. Selected discretization techniques are *Equal-Width*, *Equal-Frequency* and *K-Means*. The comparison is in Fig. 3. Our algorithm provides subgroups of higher quality for all datasets, and this no matter the applied discretization for SD-Map⋆. We infer that the information loss inherent to the attribute discretization is responsible for the poorer results obtained with SD-Map⋆. Next, we compare the run times of OSMIND and SD-Map⋆ to quantify the cost of optimality. We generate datasets - made of plant growth simulations - with sizes ranging from 10 to 10000 objects. While SD-Map⋆ and OSMIND both find the optimal subgroup in the same amount of time for small datasets, the execution time of OSMIND grows exponentially with the number of objects contrary to that of SD-Map⋆ (>40000 seconds for OSMIND vs <1 second for SD-Map⋆ with 10000 objects).

Let us now use the PCSE environment to generate 1000 random recipes. We then successively select 10, 50, 100, 200, 500 and 1000 recipes from the dataset and we observe the quality of the best subgroup returned for the quality measure q^a_{mean} when $a = 0.5$. Regarding SD-Map⋆, we use again the discretization that produces the best subgroup. Fig. 4 depicts the relative quality of the best subgroup returned by each algorithm for different dataset sizes. With smaller datasets, SD-Map⋆ finds the optimal subgroup despite the use of discretization. However, as datasets get larger, SD-Map⋆ returns consistently 10% to 25%

Table 3. Comparison: Closure on the positives (COTP) vs Normal closure (NC) and Tight improved (TI) vs Tight base (TB). "-" means execution time >72h.

Dataset	a	COTP	NC	Gain (÷)	TI	TB	Gain (%)
Bolt	0.5	25	118	4.7	0.0062	0.0078	20.5
	1	16	299	19	0.0042	0.0055	23.6
Basketball	0.5	143037	3014506	21	80.5	104	22.6
	1	42548	1121798	26	30.5	39.3	22.4
Airport	0.5	387	12042	35	0.17	0.19	10.5
	1	57	10055	176	0.033	0.037	10.8
Body Temp	0.5	795	1199	1.5	0.53	0.73	27.4
	1	570	865	1.5	0.47	0.53	11.3
Pollution	0.5	100776	-	-	23.9	25	4.4
	1	1289	41662411	32321	0.376	0.408	7.8
RecipesA	0.5	18258	430105	24	8.25	9,84	16.1
	1	1147	24431	21	0.72	0,82	12.2
RecipesB	0.5	324116	854873	2.6	1666	2223	25
	1	5261	17848	3.4	45.8	64,3	28.8

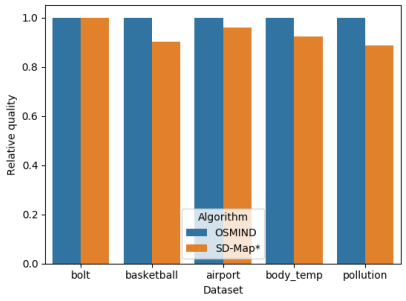


Fig. 3. Comparison of the best subgroup quality.

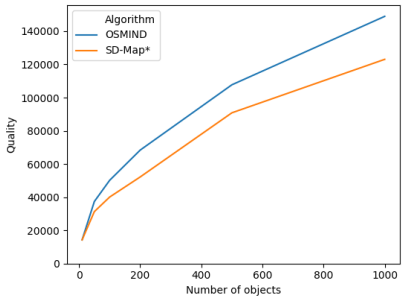


Fig. 4. Comparison of the best subgroup quality w.r.t. number of objects.

worse results. Another important qualitative aspect concerns the descriptions of the optimal subgroups found by `OSMIND` and `SD-Map*`. Table 4 depicts these descriptions for dataset RecipesA. Besides the higher quality of the subgroup returned by `OSMIND`, its description also enables to extract much more information than the description obtained with `SD-Map*`. In fact, where `SD-Map*` only offers a strong restriction on attribute $Irrad^{P2}$, `OSMIND` provides actionable information on 5 of the 9 considered attributes.

Let us finally introduce our use case on urban farm recipe optimization that is studied in [17]. We do not have access to real farming data yet but we found a way to support our application scenario thanks to the inexpensive experiments enabled by the simulator PCSE. In an urban farm, plants grow in a controlled environment (e.g., temperature, CO_2 concentration, etc). A growth recipe is the set of development conditions of a plant throughout its growth. In the absence of failure, recipe instructions are followed and an optimization objective can concern the yield at the end of the growth cycle. Table 2 features examples of growth recipes and we can simulate the execution of recipes through the use of the PCSE environment by setting the characteristics (e.g., the climate) of the different stages. We use this simulator to generate 30 recipes with random growth conditions. We focus on 3 variables that set the amount of solar irradiation, wind and rain. The plant growth is split in 3 stages of equal length. We can first check that `OSMIND` enables the discovery of a subgroup maximizing the yield. Next,

Table 4. Comparison between descriptions of: the overall dataset (DS), the optimal subgroup returned by `OSMIND`, the optimal subgroup returned by `SD-Map*`. "-" means no restriction on the attribute compared to DS, Q and S denote respectively the quality and size of the subgroup.

Subgroup	$Rain^{P1}$	$Irrad^{P1}$	$Wind^{P1}$	$Rain^{P2}$	$Irrad^{P2}$	$Wind^{P2}$	$Rain^{P3}$	$Irrad^{P3}$	$Wind^{P3}$	Q	S
DS	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	0	100
OSMIND	-	[4428,23285]	[0,27]	[8,40]	[16428,25000]	-	[2,40]	-	-	50147	26
SD-Map*	-	-	-	-	[19000,25000]	-	-	-	-	40069	31

we validate the interpretability and actionability of the return results. Table 5 features a comparison between the interval pattern of the overall dataset and that of the optimal subgroup returned by OSMIND. These results illustrate the capacity of OSMIND to discover a recipe subgroup with optimal yield (17819 vs 7256). We can use the description of the optimal subgroup as a new recipe that will lead to higher yields. The optimal interval pattern is easily interpretable and it supports the extraction of non-trivial knowledge. As an example, during the first stage of the growth cycle, the amount of solar irradiation (Irrad^{P1}) that plants undergo seems to have no impact on the optimization of the yield. This can be inferred from the weak restriction applied on the interval of values taken by Irrad^{P1}. Domain knowledge confirms: the capacity of plant light absorption is severely limited during the first stage of the growth cycle meaning that the growth cost could be cut down while keeping the same yield by restricting the amount of light used during the beginning of the plant growth.

Table 5. OSMIND results. Interval patterns of the overall dataset (DS) and the optimal subgroup returned (OS), and average Yield (Y) of recipes for each subgroup.

Subgroup	Rain^{P1}	Irrad^{P1}	Wind^{P1}	Rain^{P2}	Irrad^{P2}	Wind^{P2}	Rain^{P3}	Irrad^{P3}	Wind^{P3}	Y
DS	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	[0,40]	[1000,25000]	[0,30]	7256
OS	[0,40]	[2714,23285]	[0,21]	[8,37]	[16428,25000]	[0,23]	[2,40]	[6142,25000]	[0,27]	17819

6 Conclusion

We investigate the optimal subgroup discovery with respect to a quality measure in purely numerical data. We motivated the reasons why existing methods achieve suboptimal results by requiring a discretization of numerical attributes. The OSMIND algorithm enables optimal subgroup discovery without such a loss of information. The empirical evaluation has illustrated the added-value and the exploitability of the OSMIND algorithm when compared to the reference algorithm SD-Map$\star$. From an applicative perspective, our future work concerns the design of optimization methods for urban farms that push much further the applicaion case that was just sketched here. From an algorithmic perspective, our future work concerns the enhancement of OSMIND scalability for high-dimensional datasets. Moreover, it would be interesting to investigate how to exploit some sequential covering techniques for computing not only an optimal subgroup but a collection of non-redundant optimal subgroups.

Acknowledgment. Our research is partially funded by the French FUI programme (project DUF 4.0, 2018-2021).

References

1. Atzmueller, M., Puppe, F.: SD-Map – a fast algorithm for exhaustive subgroup discovery. In: Proceedings PKDD. pp. 6–17 (2006)

2. Aumann, Y., Lindell, Y.: A statistical theory for quantitative association rules. In: Proceedings ACM SIGKDD. pp. 261–270 (1999)
3. Belfodil, A., Belfodil, A., Kaytoue, M.: Anytime subgroup discovery in numerical domains with guarantees. In: Proceedings ECML/PKDD (2). pp. 500–516 (2018)
4. Boley, M., Grosskreutz, H.: Non-redundant subgroup discovery using a closure system. In: Proceedings ECML/PKDD (1). pp. 179–194 (2009)
5. Bosc, G., Boulicaut, J.F., Raïssi, C., Kaytoue, M.: Anytime discovery of a diverse set of patterns with monte carlo tree search. Data Min. Knowl. Discov. **32**, 604–650 (2017)
6. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings IJCAI. pp. 1022–1029 (1993)
7. Ganter, B., Wille, R.: Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (1998)
8. Garcia, S., Luengo, J., Saez, J.A., Lopez, V., Herrera, F.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Trans. on Knowl. and Data Eng. **25**(4), 734–750 (2013)
9. Garriga, G.C., Kralj, P., Lavrač, N.: Closed sets for labeled data. J. Mach. Learn. Res. **9**, 559–580 (2008)
10. Grosskreutz, H., Paurat, D.: Fast and memory-efficient discovery of the top-k relevant subgroups in a reduced candidate space. In: Proceedings ECML/PKDD (1). pp. 533–548 (2011)
11. Grosskreutz, H., Rüping, S.: On subgroup discovery in numerical domains. Data Min. Knowl. Discov. **19**(2), 210–226 (2009)
12. Grosskreutz, H., Rüping, S., Wrobel, S.: Tight optimistic estimates for fast subgroup discovery. In: Proceedings ECML/PKDD (1). pp. 440–456 (2008)
13. Kaytoue, M., Kuznetsov, S.O., Napoli, A.: Revisiting numerical pattern mining with formal concept analysis. In: Proceedings IJCAI. pp. 1342–1347 (2011)
14. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271 (1996)
15. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast exhaustive subgroup discovery with numerical target concepts. Data Min. Knowl. Discov. **30**(3), 711–762 (2016)
16. Mampaey, M., Nijssen, S., Feelders, A., Knobbe, A.: Efficient algorithms for finding richer subgroup descriptions in numeric and nominal data. In: Proceedings IEEE ICDM. pp. 499–508 (2012)
17. Millot, A., Cazabet, R., Boulicaut, J.F.: Actionable subgroup discovery and urban farm optimization. In: Proceedings IDA (2020), 12 pages, In Press.
18. Moreland, K., Truemper, K.: Discretization of target attributes for subgroup discovery. In: Proceedings MLDM. pp. 44–52 (2009)
19. Nguyen, H.V., Vreeken, J.: Flexibly mining better subgroups. In: Proceedings SIAM SDM. pp. 585–593 (2016)
20. Soulet, A., Crémilleux, B., Rioult, F.: Condensed representation of emerging patterns. In: Proceedings PAKDD. pp. 127–132 (2004)
21. Webb, G.I.: Discovering associations with numeric variables. In: Proceedings ACM SIGKDD. pp. 383–388 (2001)
22. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings PKDD. pp. 78–87 (1997)