A Clustering-Inspired Quality Measure for Exceptional Preferences Mining—Design Choices and Consequences

Ruben Franciscus Adrianus Verhaegh, Jacco Johannes Egbert Kiezebrink, Frank Nusteling, Arnaud Wander André Rio, Márton Bendegúz Bendicsek, Wouter Duivesteijn(✉), and Rianne Margaretha Schouten(✉)

Eindhoven University of Technology, Eindhoven, the Netherlands
{r.f.a.verhaegh,j.j.e.kiezebrink,f.nusteling,a.w.a.rio, m.b.bendicsek}@student.tue.nl, {w.duivesteijn,r.m.schouten}@tue.nl

Abstract. Exceptional Preferences Mining (EPM) combines the research fields of Preference Learning and Exceptional Model Mining. It is a local pattern mining task, where we try to find coherent subgroups of the dataset featuring unusual preferences between a fixed set of labels. We introduce a new quality measure for Exceptional Preferences Mining, inspired by concepts from Clustering. On top of that, we draw conclusions on two design choices that must necessarily be made whenever one defines a quality measure for any version of Exceptional Model Mining: on the one hand, exceptional behavior is easily (spuriously) found in tiny subgroups, so what is the best way to compensate for that; on the other hand, when gauging exceptionality of a subgroup's behavior, what does one use as reference for the normal behavior? We find that the choice of correction factor not only influences the subgroup size but it also effects the presumed exceptionality of found subgroups. The entropy function allows for detecting exceptional subgroups of a meaningful size, both when a candidate subgroup is evaluated against its complement and against the entire dataset.

Keywords: Exceptional preferences mining · Label ranking · Exceptional model mining · Preference learning · Pattern mining

1 Introduction

Exceptional Preferences Mining (EPM) [15,16] combines the two research fields of Preference Learning (PL) [5] and Exceptional Model Mining (EMM) [4,12]. In PL, rather than predicting the relevance of individual labels for records of the dataset, the focus lies on learning whether a record of the dataset prefers a label over another. Hence, PL is mostly concerned with analyzing how labels relate to each other, rather than the individual expression of a single label. A subfield

P. Pascal and D. Ienco (Eds.): DS 2022, LNAI 13601, pp. 429–444, 2022.
https://doi.org/10.1007/978-3-031-18840-4_31

of PL is Label Ranking (LR) [2,17], where one tries to learn a preference order (ranking) on a set of labels. This is the part of PL that is of specific concern to EPM. The other research field, EMM, seeks interesting subgroups of the dataset. A subgroup is interesting if it satisfies two properties. On the one hand, subgroups must be *interpretable*: we must be able to define them in terms of few conditions on attributes of the dataset, so that we can understand and build real-life policies on them. On the other hand, subgroups must be *exceptional*: a few columns of the dataset are split off to form the target space, over which we build a model, and subgroups are interesting if their behavior in this target space is unusual. For instance, when analyzing sequential data in target space, Markov chains can capture behavior in a subgroup, and one could assess exceptionality of Markov model parameters to gauge the quality of a subgroup [18]. Within EPM, the exceptionality of a label ranking becomes the target concept of an EMM run: we find subgroups displaying unusual rankings of a set of labels.

Existing EPM quality measures [15,16] gauge exceptionality of the label ranking within a subgroup on three separate levels of granularity (discussed in more detail in Sect. 3), but they all share one trait: they only assess whether records of the subgroup behave exceptionally, but not whether there is consistency behind the measured exceptionalities. These measures neglect that exceptionality of behavior might be achieved by lumping together disparate, heterogeneous kinds of behavior (cf. [1] for a similar argument in Subgroup Discovery, correcting for dispersion). In this paper, we propose a quality measure for EPM that not only captures exceptional behavior, but additionally encourages subgroups to have homogeneous target distributions. More specifically, we propose a quality measure for EPM based on the principles of clustering, where one optimizes for low within-cluster and high between-cluster distance. Comparably, our proposed quality measure assigns a high quality value to subgroups with preference relations that are dissimilar compared to records outside the subgroup but simultaneously very similar across records inside the subgroup.

When developing a quality measure for EMM (and hence also for EPM), two design choices must be made. On the one hand, exceptional behavior is easily (spuriously) found in tiny subgroups, so one must incorporate a component in the quality measure to promote non-tiny subgroups. Typical solutions are using the entropy of the subgroup/complement split, the size of the subgroup, or the square root thereof. On the other hand, exceptional behavior cannot exist in a vacuum: behavior can only be exceptional w.r.t. a reference behavior. Typical choices are using the behavior on the entire dataset as normal behavior, or using the behavior on the subgroup's complement as reference. Crucially, for both these design choices, very little evidence exists on what the right choice would be. In this paper, we show that the choice of correction factor not only influences the subgroup size but it also effects the presumed exceptionality of found subgroups, and we further demonstrate differences in outcomes under different reference behaviors in the context of EPM.

1.1 Main Contributions

The main contributions of this paper are:

1. a new quality measure for EPM that allows for the finding of exceptional and coherent subgroups in both descriptive and target space;
2. an exploration of the effect of subgroup size correction functions on the exceptionality of the found subgroups;
3. a demonstration of how outcomes differ depending on whether a subgroup is evaluated against the global model or against its complement.

2 Preliminaries

Exceptional Preferences Mining (EPM) [15,16] is a mix of Preference Learning (PL) [5] on the one hand and Exceptional Model Mining (EMM) [4,12] on the other hand. It combines the task of "learning to rank" [5, p. 3] with the task of identifying subgroups in a dataset that behave exceptionally. Specifically, EPM focuses on Label Ranking (LR) [2,17], a type of problem in PL that aims to map instances to rankings over a predefined set of labels, or classes. One can consider LR to be a variant of the conventional classification problem, but instead of assigning a case to a specific class, LR aims to assign a complete order of labels.

Assume a dataset Ω, which is a bag of N records $r \in \Omega$ of the form

$$r = (a_1, \ldots, a_k, t_1, \ldots, t_\ell)$$

where k and ℓ are positive integers. Target attributes $t_1, \ldots, t_\ell$ contain values associated with ℓ unique labels or classes from the set $\mathcal{L} = \{\lambda_1, \ldots, \lambda_\ell\}$. Thus, t_1 contains values associated with label λ_1, t_2 contains values associated with label λ_2, etcetera. The exact meaning of the values depends on the application domain. For instance, in a classification problem, t_v can be the probability that a record r belongs to class $\lambda_v \in \mathcal{L}$. Alternatively, in Sect. 6 of this paper, we analyze the Dutch parliament elections in 2021 and consider record $r \in \Omega$ to be a municipality; attributes $t_1, \ldots, t_\ell$ contain the number of votes for ℓ distinct political parties.

2.1 Order Relations

We are interested in the ordering of the political parties by the number of votes. The idea is to construct an ordering of the associated labels such that label λ_v precedes λ_w when $t_v > t_w$, $v \neq w$ and $1 \leq v, w \leq \ell$. Here, we consider total order relations $\succ$ on $\mathcal{L}$, which means that label λ_v cannot have the same position as λ_w. In other words, the ordering is a ranking and $\lambda_v \succ \lambda_w$ not only means that λ_v precedes λ_w but also that it is preferred over λ_w. Depending on the application, the user can decide what total order should be assigned to labels with equal values. In the case of Dutch elections, political parties with an equal number of votes will be ranked based on their position on the voting list.

Formally, a total order $\succ$ is a permutation π of the set $\{1, 2, \ldots, \ell\}$ such that $\pi(v)$ is the position of label λ_v in the order. For instance, if we consider the total order $\lambda_4 \succ \lambda_1 \succ \lambda_3 \succ \lambda_2$ for $\ell = 4$, $\pi = (2, 4, 3, 1)$.

Table 1. Example toy datasets: the shared descriptor space, and separate target spaces for SD, EMM, and EPM.

Attribute name	a_1	a_2	a_3	a_4	...	a_k
Meaning	Name	Legs	Swims?	Flies?	...	Fluffy?
r^1	Cat	4	no	no	...	a bit
r^2	Fish	0	yes	no	...	no
r^3	Owl	2	no	yes	...	no
r^4	Sheep	4	no	no	...	very yes
r^5	Snail	0	no	no	...	no

(a) Descriptor space

t_1
Friendly
no
yes
no
yes
yes

(b) SD

t_1	t_2	...	t_m
Length	Weight	...	Life span
46	4 500	...	15
10	227	...	12
41	1 585	...	8
1 500	95 000	...	11
2	6	...	6

(c) EMM

$\pi(1)$	$\pi(2)$	$\pi(3)$
Grass rank	Bread rank	Meat rank
2	3	1
2	1	3
2	3	1
1	2	3
1	2	3

(d) EPM ($\ell = 3$)

2.2 Local Pattern Mining Methods: SD, EMM, and EPM

In the setting of both LR and EPM, preferences on $\mathcal{L}$ are associated with particular (groups of) dataset records through a set of features or attributes. In EMM and EPM terms, these features are *descriptive* attributes, or descriptors. Attributes $a_1, \ldots, a_k$ are these descriptors. The task of Local Pattern Mining methods [7,13] is to find subgroups of the dataset, defined as a conjunction of conditions on a few descriptors. Subgroup Discovery (SD) [8,10,20] seeks subgroups displaying an unusual distribution of a single target attribute, Exceptional Model Mining (EMM) [4,12] seeks subgroups displaying an unusual interaction between multiple target attributes, and Exceptional Preferences Mining (EPM) [15,16] seeks subgroups where this interaction is exceptional preference relations. Hence, EMM can be seen as the multitarget generalization of SD, and EPM can be seen as a specific instantiation of the generic EMM framework.

Table 1 displays a toy dataset of some animals in a zoo. SD, EMM, and EPM all share the descriptor space of Table 1a; any target space from Tables 1b, c, and d can be appended. Combining Tables 1a and b, SD would find that the subgroup "flies? = no" has a 75% share of "friendly = yes", while this share is 60% in the overall population. Combining Tables 1a and c, EMM would find subgroups with an unusual interaction between the m targets (for example, exceptional regression coefficients of length and weight while predicting life span, when using the EMM model class from [3]). Combining Tables 1a and d, EPM would find that the subgroup "Legs $\leq$ 1" always ranks meat last.

2.3 Definitions

The task of EPM is to identify subgroups in the dataset with exceptional preferences. The subgroups are defined by a description over the collective domain of descriptive attributes. Formally, a description is a function $D : \mathcal{A} \mapsto \{0, 1\}$, and a record r^i is covered by description D if and only if $D(a_1^i, ..., a_k^i) = 1$.

Definition 1. The subgroup corresponding to a description D is the bag of records $S_D \in \Omega$ that D covers: $S_D = \{r^i \in \Omega \mid D(a_1^i, \ldots, a_k^i) = 1\}$.

We denote the number of records in a subgroup S with n. Every subgroup has a complement $S^C = \Omega \setminus S$ which contains all $n^C = N - n$ records not in S. Whether a subgroup has exceptional preferences is evaluated with a *quality measure* (QM):

Definition 2. Given a description language $\mathcal{D}$ governing which subgroups can be formulated on a given dataset, a quality measure is a function $\varphi : \mathcal{D} \mapsto \mathbb{R}$.

The goal is to find the top-q subgroups with the highest quality value. It is practically impossible to investigate all candidate subgroups exhaustively since the number of candidates scales exponentially with the number of descriptive attributes. Therefore, we perform a heuristically guided search called beam search. We will further discuss beam search in Sect. 4.1 while discussing the time complexity of our approach.

3 Related Work

Local pattern mining methods have been used to understand preference relations. For instance, the Olympic ranking of countries has been studied [14] with SD. Casting German federal Bundestag election vote shares (and vote share changes between subsequent elections) within regions as preference relations, a traditional SD analysis can be performed by averaging across the ℓ parties [6]. Instead, we are interested in finding subgroups with an unusual interaction between ℓ target attributes (and therefore consider our approach to be EMM).

Existing EPM QMs [15,16] are based on preference matrices (PM). A PM $\in \{-1, 0, 1\}^{\ell \times \ell}$ is a square matrix that for each pair of labels λ_v, λ_w in a ranking π evaluates whether they precede (1) or succeed (-1) each other ($\forall v, w \in \{1, \ldots, \ell\}$). PMs of individual records can be averaged, which allows for the comparison of matrix M^D, the PM for the entire dataset, with M^S, the PM for the subgroup. Denoting the difference between matrices M^D and M^S with L^S, [15,16] propose three quality measures, for exceptionality on three distinct levels of behavioral granularity:

$$\varphi_{\text{norm}} = \sqrt{n/N} \cdot \sqrt{\sum_{v=1}^{\ell} \sum_{w=1}^{\ell} L^S(v,w)^2} \tag{1}$$

$$\varphi_{\text{labelwise}} = \sqrt{n/N} \cdot \max_{v=1,\ldots,\ell} \frac{1}{(\ell-1)} \sum_{w=1}^{\ell} L^S(v,w) \tag{2}$$

$$\varphi_{\text{pairwise}} = \sqrt{n/N} \cdot \max_{v,w=1,\ldots,\ell} L^S(v,w) \tag{3}$$

The first QM, φ_{norm}, takes the Frobenius norm of L^S to search for preference deviations that occur spread out across the entire difference matrix. Zooming in, $\varphi_{\text{labelwise}}$ evaluates whether there is one particular label λ_v that ranks substantially different in the subgroup, ignoring interactions between other labels. Zooming in even further, $\varphi_{\text{pairwise}}$ studies pairwise preferences [9], evaluating whether any pair of labels interacts unusually in the subgroup. All three QMs compare the PM of the subgroup with the PM of the entire dataset, and share the choice for subgroup size correction factor:

$$\xi_{\text{sqrt}} = \sqrt{n/N}. \tag{4}$$

In developing our quality measure we will borrow principles from clustering. EMM is a local pattern mining technique whereas clustering is a global analysis task, partitioning all records into homogeneous clusters. In EMM, subgroups have an interpretable description, and records may be assigned to any number of subgroups. Methods on the crossroads of local and global pattern mining have been proposed, such as Predictive Clustering Rules (PCR) [21], SD with a classification rule learning algorithm (CN2-SD) [11], Cluster Grouping (CG) [22] and Multi-Response Subgroup Discovery (MR-SD) [19]. Although our quality measure is inspired by principles in clustering, our method is a purely local one.

4 Proposed Method: A Clustering-Based Quality Measure

We propose to perform EPM using the following clustering-based quality measure. Given a subgroup S and its complement S^C, let π^i denote the ranking of labels in the i^{th} record of S, and let π^j denote the ranking in the j^{th} record of S^C, where $1 \leq i \leq n$ and $1 \leq j \leq n^C$. We seek subgroups of records with exceptional label preferences. Those subgroups should have rankings dissimilar from the rankings in its complement. We define this notion of inter-subgroup distance as

$$\alpha_{\text{compl}} = \frac{1}{n \cdot n^C} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n^C} d(\pi^i, \pi^j), \tag{5}$$

where $d(\cdot,\cdot)$ is some distance metric between the two rankings.

In addition, we want the cases in the subgroup to have similar rankings (i.e. to have small distance to one another), because coherent and homogeneous

subgroups are 1) easier to interpret and 2) more practically relevant than heterogeneous subgroups. We define this notion of intra-subgroup distance as

$$\beta = \frac{1}{n \cdot (n-1)} \cdot \sum_{h=1}^{n} \sum_{i=1}^{n} d(\pi^h, \pi^i). \tag{6}$$

Next, we divide the inter-subgroup distance α by the intra-subgroup distance β, which results in the following quality measure,

$$\varphi_{\text{clus}} = \frac{\alpha}{\beta + 1} \cdot \xi, \tag{7}$$

where ξ is a function that corrects for the subgroup size. We add 1 to the denominator to account for perfect homogeneous subgroups (where $\beta = 0$).

Quality measure φ_{clus} is expected to give a high value when the subgroup is homogeneous (β small), when the subgroup's rankings are different from those in its complement (α large) or, ideally, both. Hence, our proposed quality measure is generic. Simultaneously, the distance function $d(\cdot, \cdot)$ can be specified by the user, which allows for searching subgroups with specific ranking deviations.

If one is interested in comparing a subgroup with the average ranking in the entire dataset, Eq. (5) can easily be adapted as follows,

$$\alpha_{\text{average}} = \frac{1}{n} \sum_{i=1}^{n} d(\pi^i, \pi^D), \tag{8}$$

where π^D is the label ranking when all N data records are taken into account. In this scenario, β does not change: we still aim to find coherent subgroups with exceptional label rankings; only the reference behavior has changed.

4.1 Time Complexity

To traverse the space of candidate subgroups, we apply beam search, a commonly used algorithm that is flexible in handling descriptive attributes of binary, categorical, and/or numerical type [4]. The algorithm performs a level-wise search of d levels, where the first level evaluates candidate subgroups with descriptions based on 1 descriptor and each subsequent level refines the descriptions of the top-w subgroups. The time complexity of beam search for EMM [4] is given by

$$\mathcal{O}(dwkE(c + \mathcal{M}(N, \ell) + \log(wq))), \tag{9}$$

where E is the worst-case number of categories (binary and numerical attributes are refined faster), c refers to the complexity of comparing the model in the subgroup against another model, $\mathcal{M}(N, \ell)$ is the cost of learning a model on N records and ℓ targets and d, w, k, and q are as described before (cf. [4, Section 4.2.1] for more details).

To evaluate the exceptionality of a candidate subgroup with quality measure φ_{clus}, α_{compl} requires $n \cdot n^C$ comparisons, α_{average} requires n comparisons and β

requires $n \cdot (n-1)$ comparisons. The time complexity of one comparison depends on the number of target attributes ℓ. That means that the time complexity of calculating φ_{clus} scales quadratically: $\mathcal{O}(\ell(n \cdot (n-1) + n \cdot n^C)) = \mathcal{O}(N^2 \cdot \ell)$ (since n and n^C are both $\mathcal{O}(N)$). The effect of c is already incorporated here.

The original EPM QMs [15,16] have a different time complexity. Calculating a PM costs $\mathcal{O}(\ell^2)$ per record; an average PM over n records then has a complexity of $\mathcal{O}(n\ell^2)$. In Sect. 5, we will further analyze these run times with synthetic data. Section 6 evaluates the performance of our proposed quality measure on real-world data.

4.2 Qualitative Differences Between φ_{clus} and Existing QMs

The added value of the quality measure is that it finds interesting results based on the distance between the sum of the permutations of the subgroup and the complement of the subgroup. Therefore, this quality measure should excel in finding those subgroups where the general ranking of the target variables differs greatly. Where previous work uses a general mean norm quality measure to find subgroups for label ranking [6], φ_{clus} seems intuitively very similar to a norm-based quality measure. It is different in that it tries to find subgroups based on the deviation from the overall mean of the permutations of the labels.

Existing work introduces different approaches in order to find subgroups for label ranking. Within EPM, preference matrices [16] are used; beyond pattern mining, a meta learning technique to reduce label ranking to binary classification was proposed [2]. Both these papers rely on preference matrices: label ranks are transformed to an interval $[0, 1]$ by averaging preferences of label pairs, thus accumulating them to matrices, one representing the dataset (M_D) and another representing the subgroup (M_S) [16]. Our algorithm, on the other hand, calculates the average distance of a label in the subgroup compared to those within the subgroup S (β) and to those in the complement subgroup S^C (α). The quality measures presented in the studied literature all have clear use cases as mentioned in Sect. 3, while our measure aims to be more generic.

The approach of the quality measure created in this paper is different from all above-mentioned ones, thus could yield different interesting results. Besides this, φ_{clus} should be robust with respect to variations in dataset metacharacteristics that theoretically ought not to negatively affect the outcome of an Exceptional Preferences Mining run. More specifically, the number of rows in the dataset will likely not influence the quality measure as the similarities are normalized. An increase in the number of target variables will likely make finding subgroups more stable: more target variables will reduce the opportunity for sudden peaks in the distance function. The number of descriptive attributes in a dataset almost always affects local pattern mining techniques such as EPM: an increase in the number of descriptors exponentially inflates the search space, making interesting subgroups harder to find. The expectation is that this will be no different for this quality measure.

5 Synthetic Data Experiment

We generate data with $N \in \{100, 500, 1000\}$ records. Each of these records can be described by $k \in \{2, 8, 32\}$ binary descriptors, which are independently sampled from a binomial distribution $a_h \sim \text{Bin}(N, p)$ with $p = 0.5$ for all $1 \leq h \leq k$. For the sake of simplicity and consistency, we let the true subgroup cover records where $a_1 = 1 \wedge a_2 = 1$, resulting in subgroups with size $n = \frac{1}{4}N$.

Each record has a ranking based on $\ell \in \{2, 8, 32\}$ target attributes. Since we want our synthetic data to resemble a real-world scenario as much as possible, we first analyze the average ranking of the $\ell = 37$ political parties in the real-world dataset (see Sect. 6). There, a party with rank $v+1$ has about 0.7 times as many votes as the party with rank v, for all $1 \leq v \leq \ell$. The variance of the number of votes over the records had an average ratio with the number of votes of 0.03. For the synthetic dataset, we therefore draw ℓ target attributes from a normal distribution $t_v \sim \mathcal{N}(\mu_v, \sigma_v^2)$ with mean $\mu_v = 0.7^{(v-1)}$ and variance $\sigma_v^2 = 0.03\mu_v$. Given the number of votes per party, a ranking π per record is obtained as per Sect. 2. Because of random sampling, $\pi(v)$ may or may not have value v, but on average the ranking in the entire dataset will be $\pi^D = (1, 2, \ldots, \ell)$.

We experiment with three types of subgroups (N.B.: every dataset contains one true subgroup, whose type is a simulation parameter):

reversed: we invert the values of the target attributes; the values of t_1 are swapped with the values of t_ℓ, the values of t_2 are swapped with $t_{\ell-1}$, etc. On average, this will result in $\pi_{\text{rev}} = (\ell, \ell-1, \ldots, 2, 1)$.

pairwise-swapped: we swap each consecutive pair of attributes; the values of t_1 are swapped with the values of t_2, the values of t_3 are swapped with t_4, etc. This will result in $\pi_{\text{pair}} = (2, 1, 4, 3, \ldots, \ell, \ell-1)$ for even values of ℓ.

last-to-first: here, no matter the values of attribute t_ℓ, we put λ_ℓ at rank 1, resulting in $\pi_{\text{ltf}} = (2, 3, \ldots, \ell-1, \ell, 1)$.

Note that because we generate the entire dataset first, and then replace the target values of the records covered by the true subgroup definition, $\pi^{SG^C} = \pi^D$.

We evaluate the performance of φ_{clus} with $\alpha = \alpha_{\text{compl}}$, and experiment with three types of subgroup size corrections: ξ_{sqrt} as given in Eq. (4), the entropy function as proposed for EMM by [12],

$$\xi_{\text{entropy}} = -\frac{n}{N}\log\frac{n}{N} - \frac{n^C}{N}\log\frac{n^C}{N}, \tag{10}$$

and no correction: $\xi_{\text{none}} = 1$. The three ways of correction are chosen such that they have opposite objectives: ξ_{sqrt} prefers larger subgroups, ξ_{entropy} prefers a 50/50 split of the dataset and ξ_{none} guides the search towards small subgroups. We run beam search with $w = 20$, $d = 3$ and evaluate whether or not the true subgroup is present in the top-q subgroups with $q = 10$. For every combination of parameters, we run the experiment $nreps = 5$ times and report the proportion of true subgroups not appearing in the top-q result list, the average rank of the true subgroups in that result list and the average run time. See https://github.com/bendicsekb/data_mining_election for all source code and results.

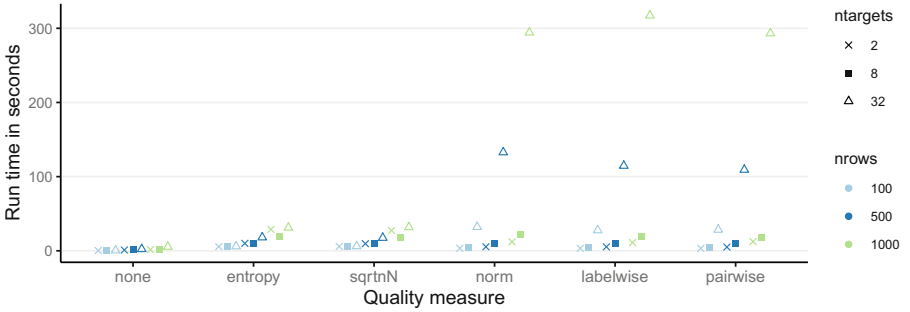

Fig. 1. Run time in seconds of the beam search algorithm for six quality measures, varying dataset size and varying number of target attributes. (Color figure online)

5.1 Results

For the reversed subgroup type and the last-to-first subgroup type, all six QMs ($\varphi_{\text{norm}}, \varphi_{\text{labelwise}}, \varphi_{\text{pairwise}}$, and φ_{clus} with three variants of ξ) find the ground truth subgroup in 100% of the cases at the first position in the result list. Quality measure φ_{clus} finds the pairwise-swapped subgroup type under all simulation conditions when $\xi = \xi_{\text{none}}$ (when no subgroup size correction is applied). For ξ_{entropy} and ξ_{sqrt}, the true subgroup cannot be found when both k and ℓ are 32. Like φ_{clus} with ξ_{none}, quality measures φ_{norm} and $\varphi_{\text{labelwise}}$ find the pairwise-swapped subgroup under all simulation conditions. When $N = 100$, $\varphi_{\text{labelwise}}$ has difficulty when the number of descriptors k is too large relative to the number of targets ℓ (8 vs. 2, 32 vs. 8 and 32 vs. 32). The problem disappears when N increases to 500 or 1000.

Figure 1 presents the run times in seconds for all six quality measures for the reversed subgroup type. Conclusions are similar for the pairwise-swapped and last-to-first subgroup type. As discussed in Sect. 4.1, we see that φ_{clus} scales with the number of rows N while the EPM QMs scale with the number of targets ℓ.

6 Real-World Data Experiment

We analyze data from the 2021 Dutch general election, publicly available at https://www.verkiezingsuitslagen.nl/verkiezingen/detail/TK20210317. The dataset contains the number of votes for $\ell = 37$ political parties in $N = 351$ municipalities. We add information about socio-economic characteristics of the municipalities such as the number of citizens, gender balance, age distribution, migration background, number of companies, how many ducks go for slaughter (proxy for rurality), total road length, and much more. That dataset is made available by Statistics Netherlands (https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=70072ned&_theme=237). Consequently, we have $k = 83$ numerical descriptors.

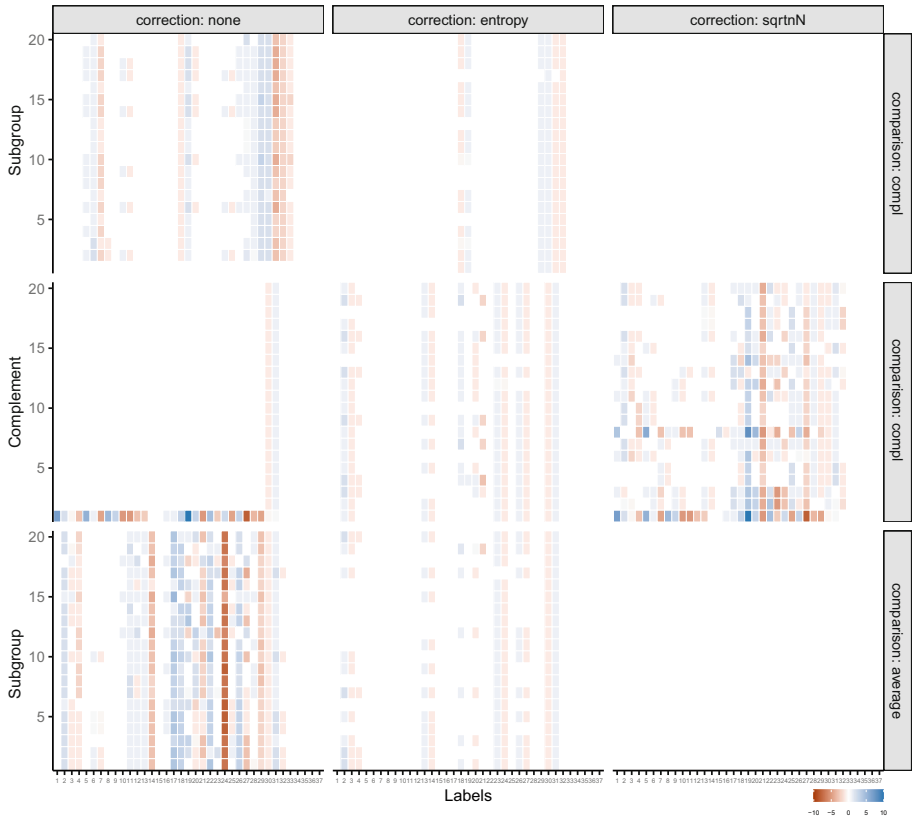

Fig. 2. Difference in rank between a group and the average dataset ranking π^D, obtained with quality measure φ_{clus} (red: higher rank in group, blue: lower rank in group, white: no difference). (Color figure online)

The goal is to find coherent subgroups of municipalities with an exceptional ranking of political parties. We evaluate the performance of φ_{clus} in six scenarios, combining the two comparisons with reference models α_{compl} and α_{average} with three types of subgroup size correction ξ_{none}, ξ_{entropy}, and ξ_{sqrt}. As distance function $d(\cdot, \cdot)$ within φ_{clus} we employ the Euclidean norm. The beam search is run with parameters $w = 30$, $d = 3$, and $q = 20$, and minimum subgroup size constraint of $c_{\text{size}} = 10\%$.

Figure 2 presents results. Each of the 9 panels shows the difference in label ranking between a subgroup and the average dataset ranking π^D ($q = 20$ by $\ell = 37$). Here, red indicates that a political party has a higher rank in the subgroup (more votes, moved to the left), blue represents a lower rank in the subgroup (fewer votes, moved to the right) and white means that there is no difference. The three columns in Fig. 2 correspond to the three types of subgroup size correction. The three rows correspond to reference models α_{compl} (row 1 and

2, Sect. 6.1) and α_{average} (row 3, Sect. 6.2). The panels in the first two rows use the same beam search results, but the top row gives the difference in ranking for the discovered subgroups whereas the second row shows the results for the complements of the discovered subgroups.

6.1 Comparing Against the Complement

When no subgroup size correction is applied, comparing candidate subgroups with their complements results in subgroups with quite some exceptional preferences (top left panel in Fig. 2). For instance, in the second subgroup, label λ_5 and λ_6 have moved to the right, while label λ_7 and λ_8 have moved to the left (the first subgroup is an all-white subgroup). These labels correspond to three relatively left-leaning parties (SP, PvdA, GroenLinks). Label λ_8 corresponds to FvD, a very right-leaning party. The subgroup covers municipalities with *Green pressure* $\geq$ *37.2%* $\wedge$ *Surinam migration background* $\geq$ *0.8%* $\wedge$ *Any-non western migration background* $\geq$ *9.6%*. Here, *green pressure* refers to the ratio between the number of people aged 0 to 20 and the number of people aged 20 to 65. Municipalities with a high green pressure skew younger. Our results indicate that younger citizens, or their parents, vote more extremely on the electoral spectrum than older citizens.

In the center left panel, we see that the complement of this subgroup has a label ranking that is similar to the average dataset ranking π^D, except for labels λ_{30} and λ_{31}. For all subgroups (except for the first) found with φ_{clus} using ξ_{none} and α_{compl}, the subgroups have exceptional preferences while their complements have average preferences.

We see an opposite effect when using ξ_{sqrt} to correct for subgroup size (top right panel). Here, all $q = 20$ subgroups do not deviate from the average dataset ranking. However, the complements of these subgroups show very exceptional preferences relations (center right panel). For some of these complements, a Christian party (CU) has obtained fewer votes (see blue color for λ_{10}). That happens for instance in the complement of subgroup 4, which is described by *Dutch background* $\leq$ *92%*. Apparently, in municipalities where the percentage of citizens with a Dutch background is $> 92\%$ (the complement), people tend to vote less for this particular Christian party.

Finding contrasting results for two opposite types of subgroup size correction gives us more insight in the performance of φ_{clus} specifically and quality measures in EMM in general. Remember that φ_{clus} is designed to generate exceptional and homogeneous subgroups. Then, when the algorithm divides the dataset into two groups and compares one group with the other (e.g., the complement), it rewards the more homogeneous group and chooses that to be the subgroup.

Using ξ_{sqrt} as a correction factor and giving preference to larger subgroups, the algorithm will select subgroups with records that are close to the dataset norm; it is likely that there are more records with average ranking behavior than there are records with similar non-average behavior. Using ξ_{none}, we find exceptional subgroups. Unfortunately, the consequence is that the subgroups are very small and have a size just larger than the minimum constraint.

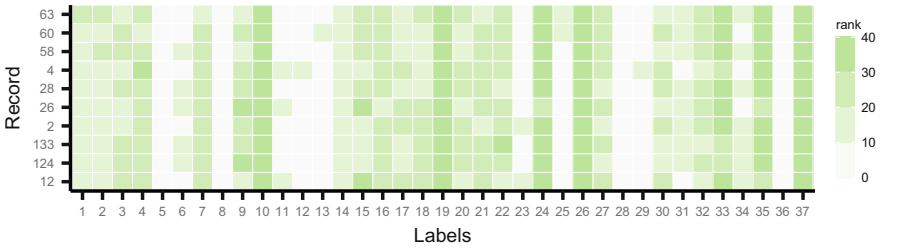

Fig. 3. Label ranking of 10 random records in the best-scoring subgroup found with φ_{clus} when comparing with the average dataset ranking and using no correction factor. (Color figure online)

The entropy function gives results that are in-between (see center top panel in Fig. 2). Although φ_{clus} still tends to select homogeneous subgroups, the preference for subgroups that contain half the number of records generates subgroups that do deviate from the norm, albeit for only a few labels. The complements of these subgroups have more exceptional preferences.

6.2 Comparing Against the Average Dataset Ranking

The bottom row of Fig. 2 presents results for φ_{clus} using α_{average} instead of α_{compl}. A clear effect of that can be seen when we use ξ_{none}. Then, the subgroups have label rankings that deviate much more from the average ranking than when we evaluate a candidate subgroup against its complement (compare left top panel with left bottom panel). In this scenario, the inter-subgroup distance as given by α_{average} is larger than α_{compl}. At the same time, the subgroups are coherent in target space and have small β, as can be seen for subgroup 1 in Fig. 3 where we present the label ranking of 10 random records in the subgroup. Although some fluctuations and differences between records exist, in general the records have similar rankings. Unfortunately, these results do not carry over to the scenario where we prefer larger subgroups (ξ_{sqrt}). Then, even though the subgroups have average ranking behavior, the intra-subgroup distance β dominates the quality value (bottom right panel in Fig. 2).

Like before, ξ_{entropy} finds an in-between solution and presents subgroups with exceptional preferences while making sure that the subgroups have a meaningful size (bottom center panel). Interestingly, the results are similar to those in the center panel, where we evaluate the complements of subgroups that are found under α_{compl}. Apparently, when using ξ_{entropy} in φ_{clus}, the reference group does not matter as much and similar exceptional subgroups are found. The difference is that these exceptional subgroups are not selected when we evaluate them against their complements, because the latter are more homogeneous and will therefore have a higher quality value.

7 Conclusions

We propose a new quality measure for Exceptional Preferences Mining (EPM) that identifies homogeneous subgroups in a dataset with unusual rankings of a set of labels. Inspired by principles from clustering, where one optimizes for low within-cluster distance or high between-cluster distance, we aim to identify subgroups with preference relations that are dissimilar compared to the rest of the dataset but very similar compared to records inside the subgroup.

As synthetic data experiments show (cf. Fig. 1), the time complexity of our quality measure scales with the number of dataset records, as opposed to existing quality measures for EPM that scale with the number of labels. Runtimes are roughly equivalent when the number of targets is 2 or 8, but our QM is substantially faster when this number is 32.

When developing a quality measure for EMM (and hence also for EPM), a correction for subgroup size should be included in order to steer the search away from tiny subgroups. Furthermore, one has to choose the reference behavior: is a candidate subgroup compared with its complement or with the average behavior? We investigate these scenarios for a real-world dataset with information about the voting behavior of municipalities in the Netherlands (cf. Fig. 2). If we compare candidate subgroups S with their complements S^C, we find that a size correction that prefers larger subgroups results in a search where the intra-subgroup distance dominates. Interestingly, exceptional ranking behavior is happening in this result set, but on the complements of the subgroups that EPM reports, which themselves display consistent but unexceptional behavior. In contrast, when we do not apply a correction for subgroup size, EPM reports subgroups are exceptional and homogeneous, but they only barely pass the minimum support constraint. To find subgroups of substantial size that also display exceptional behavior themselves, the entropy function gives the best results.

Comparing candidate subgroups with the average dataset Ω delivers subgroups with very exceptional preferences, especially when there is no correction for subgroup size. Then, the inter-subgroup distance will increase while exceptional subgroups are still coherent and homogeneous. When the entropy function is used, we again find subgroups with exceptional preferences of meaningful size. Comparing with Ω instead of S^C leads to comparable results, but the exceptional behavior is more often encompassed by the subgroups resulted by EPM instead of hidden away in their complements.

References

1. Boley, M., Goldsmith, B.R., Ghiringhelli, L.M., Vreeken, J.: Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery. Data Min. Knowl. Discov. **31**(5), 1391–1418 (2017)
2. Cheng, W., Henzgen, S., Hüllermeier, E.: Labelwise versus pairwise decomposition in label ranking. In: Proceedings of the 15th LWA Workshops: KDML, IR and FGWM, pp. 129–136 (2013)

3. Duivesteijn, W., Feelders, A., Knobbe, A.J.: Different slopes for different folks: mining for exceptional regression models with Cook's distance. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012), pp. 868–876 (2012)
4. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional model mining – supervised descriptive local pattern mining with complex target concepts. Data Min. Knowl. Disc. **30**(1), 47–98 (2016)
5. Fürnkranz, J., Hüllermeier, E.: Preference learning: an introduction. In: Fürnkranz, J., Hüllermeier, E. (eds.) Preference Learning, pp. 1–17. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14125-6_1
6. Grosskreutz, H., Boley, M., Krause-Traudes, M.: Subgroup discovery for election analysis: a case study in descriptive data mining. In: Proceedings of the 13th International Conference on Discovery Science (DS 2010), pp. 57–71 (2010)
7. Hand, D.J., Adams, N.M., Bolton, R.J. (eds.): Pattern Detection and Discovery. LNCS (LNAI), vol. 2447. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45728-3
8. Herrera, F., Carmona, C.J., González, P., Del Jesus, M.J.: An overview on subgroup discovery: foundations and applications. Knowl. Inf. Syst. **29**(3), 495–525 (2011)
9. Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artif. Intell. **172**(16–17), 1897–1916 (2008)
10. Klösgen, W.: Explora: a multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271 (1996)
11. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. J. Mach. Learn. Res. **5**, 153–188 (2004)
12. Leman, D., Feelders, A., Knobbe, A.: Exceptional model mining. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECMLPKDD 2008), pp. 1–16 (2008)
13. Morik, K., Boulicaut, J.-F., Siebes, A. (eds.): Local Pattern Detection. LNCS (LNAI), vol. 3539. Springer, Heidelberg (2005). https://doi.org/10.1007/b137601
14. Pieters, B.F., Knobbe, A., Džeroski, S.: Subgroup discovery in ranked data, with an application to gene set enrichment. In: Proceedings of the Preference Learning Workshop at Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECMLPKDD 2010), pp. 1–18 (2010)
15. de Sá, C.R., Duivesteijn, W., Azevedo, P.J., Jorge, A.M., Soares, C., Knobbe, A.J.: Discovering a taste for the unusual: exceptional models for preference mining. Mach. Learn. **107**(11), 1775–1807 (2018)
16. de Sá, C.R., Duivesteijn, W., Soares, C., Knobbe, A.: Exceptional preferences mining. In: Proceedings of the 19th International Conference on Discovery Science (DS 2016), pp. 3–18 (2016)
17. de Sá, C.R., Soares, C., Knobbe, A.: Entropy-based discretization methods for ranking data. Inf. Sci. **329**, 921–936 (2016)
18. Schouten, R.M., Bueno, M.L., Duivesteijn, W., Pechenizkiy, M.: Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. Data Min. Knowl. Disc. **36**, 379–413 (2022)
19. Umek, L., Zupan, B.: Subgroup discovery in data sets with multi-dimensional responses. Intell. Data Anal. **15**(4), 533–549 (2011)
20. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of PKDD, pp. 78–87 (1997)

21. Ženko, B., Džeroski, S., Struyf, J.: Learning predictive clustering rules. In: Proceedings of the International Workshop on Knowledge Discovery in Inductive Databases, pp. 234–250 (2005)
22. Zimmermann, A., De Raedt, L.: Cluster-grouping: from subgroup discovery to clustering. Mach. Learn. **77**(1), 125–159 (2009)