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

Diverse Subgroup Set Discovery in Multi-Label Data

Guillaume Bosc · Aimene Belfodil · Jérôme Golebiowski · Moustafa Bensafi · Jean-François Boulicaut · Marc Plantevit · Céline Robardet · Mehdi Kaytoue

Received: date / Accepted: date

Abstract Subgroup Discovery (SD) is a leading data mining technique that allows one to elicit descriptions (covering objects called subgroups) that unexpectedly occur with a class target. In presence of multi-label data, there are many applications for which one is interested in subgroups that differ from the whole dataset only on subsets of labels, and the interesting subsets of labels are not known beforehand: Typically, descriptive rules that conclude on small label sets. SD, and its extension for more complicated target concepts, Exceptional Model Mining (EMM), fail to produce such rules: They consider either the whole set of labels, each label independently, or a unique and fixed label subset chosen *a priori*. In this article, we propose to enhance EMM for considering all target subspaces (label subsets): It requires to revisit its formalization, the subgroup search space, and to propose new quality measures expressing how exceptional is a subgroup. Our quality measures consider label distributions dynamically during a heuristic search (i.e., beam search and Monte Carlo Tree Search) that outputs a diversified set of high quality subgroups, whose redundancy is controlled not only on the covered objects, but also on the considered label sets. This is shown through an extensive set of

Guillaume Bosc, Aimene Belfodil, Céline Robardet and Mehdi Kaytoue
Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
E-mail: firstname.surname@insa-lyon.fr

Marc Plantevit
Université de Lyon, CNRS, Université Lyon1, LIRIS, UMR5205, F-69621, France
E-mail: marc.plantevit@univ-lyon1.fr

Jérôme Golebiowky
Université de Nice, CNRS, Institute of Chemistry, Nice, France
E-mail: Jerome.golebiowski@unice.fr

Moustafa Bensafi
Université de Lyon, CNRS, CRNL, UMR5292, INSERM U1028, Lyon, France.
E-mail: moustafa.bensafi@cnrs.fr

experiments. Finally, we discuss an application in neurosciences which argue the actionability of the discovered subgroups.

Keywords Subgroup set discovery · Exceptional model mining · Multi-label data · Heuristic search · Monte Carlo tree search · Diversity

1 Introduction

The discovery of descriptions which distinguish a group of objects given a target (class) has been widely studied in data mining and machine learning community under several vocables (subgroup discovery, emerging patterns, contrast sets, hypotheses in formal concept analysis) [41]. We consider here descriptive rule discovery in the well-established framework of subgroup discovery (SD) [48]. Given a set of objects taking a vector of attributes (of boolean, nominal, or numerical type) as description, and a class label as target, the goal is to efficiently discover subgroups of objects for which there is a high difference between the label distribution within the group compared to the distribution within the whole dataset, e.g. considering the difference of precision with the well-known weighted relative accuracy (WRAcc) [32]. In others terms, we search for descriptive rules that conclude on a single label with possibly few errors.

In this article, we are interested in multi-label data where an object can take several class labels. For that matter, SD has been extended to a richer framework that handle more complicated target concepts, Exceptional Model Mining (EMM) [33]. The idea is the following: A model is built over the labels from the objects in the subgroup and is compared to the model of the whole dataset thanks to a quality measure (Figure 1). The more different the model, the more exceptional and interesting is the subgroup. For example, van Leeuwen and Knobbe compare the label distribution with the Weighted Kullback Leibler divergence (WKL) [46] and Duivesteijn et al. compare conditional dependence relations between the targets with Bayesian networks [18]. There exists many other models (e.g., regression, classification, target association) and associated quality measures (see [17]).

The proposed EMM instances for multi-label data consider either the whole set of labels (e.g., compared with the WKL) or each label independently (e.g., SD with the WRAcc), or finally a unique and fixed label subset chosen *a priori* (e.g. SD with the WRAcc for a label subset only). However, in presence of multi-label data, there are many applications for which one is interested in subgroups that differ from the whole dataset only on subsets of labels, and the interesting subsets of labels are not known beforehand: Typically, descriptive rules that conclude on small label sets. As an example, in the case study we report in this paper, we are interested in rules between physicochemical properties of odorant molecules (e.g. molecular weight, atom count), the attributes, and perceived smells given by Humans (e.g., fruity, apple, wood) which are the labels. There is a crucial need to discover such rules for a better understanding of the olfactory percept: According to a recent study in Science

journal, some odors can be predicted [26], but there is a few knowledge that would explain why a molecule smells an odor or another. On average, each odorant molecule is depicted by 2.33 labels among the set of 74 possible labels. Assessing a subgroup with regard to either the whole set of labels or each label independently cannot allow to characterize rules involving a subset of labels. For example, Figure 3 shows the label distribution of the best subgroup according to the WKL in our olfaction dataset: It fails at characterizing a small set of labels.

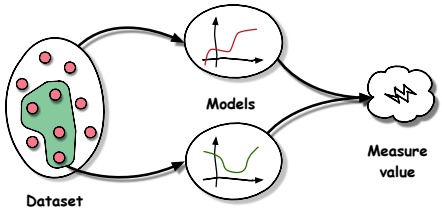


Fig. 1 A typical EMM instance.

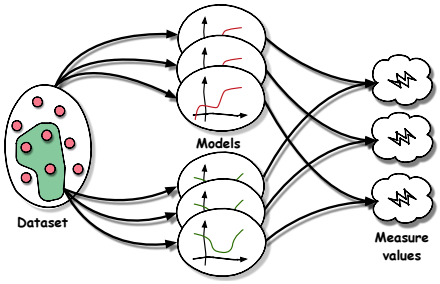


Fig. 2 EMM considering target subspaces.

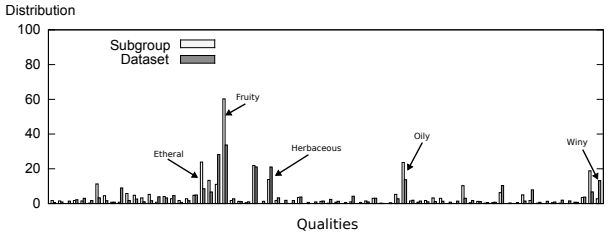


Fig. 3 Label distributions of a subgroup output by EMM.

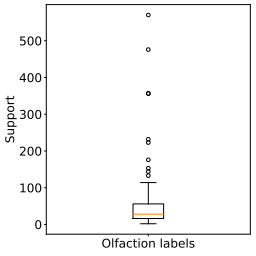


Fig. 4 Olfaction data label distribution.

The only solution for applying SD/EMM and successfully extract subgroups characterized by multi-label is to transform the data. Indeed, when facing a multi-label dataset, either for a supervised or unsupervised task, the most popular approaches apply standard techniques after a data transformation: Binary Relevance (BR) creates a dataset for each different label (objects are kept, target becomes binary) while Label Powerset (LP) creates a dataset for each set of labels that exists in the data set [43]. In our settings, BR looses label dependencies (e.g. fruity and apple), but applying LP before classical SD/EMM is totally relevant. It requires however to consider as many datasets as label subsets in the data, $1,800$ combinations in our olfaction dataset which

could be restricted to 400 judged as interesting by the neuroscientists who created the dataset. This comes with an explosion of computational times and explosion of rules returned to the expert. As explained hereafter, SD/EMM is nowadays shifting towards the heuristic search of a diverse set of subgroups that is large enough to cover interesting label sets, and small enough to be analyzed by an expert. As such, as a first contribution, we properly formalize a novel EMM instance to fully take into account multi-label data and all label subsets called *target subspaces* (Figure 2). This comes with a more complex search space that needs even more to be handled with heuristic searches.

Indeed, it has been now widely admitted that an exhaustive search is impossible in general, even with efficient pruning techniques [46,17,6]: The subgroup search space is large and it is difficult to use the properties of the measures to prune the search space. Accordingly, heuristic approaches are used, mainly based on beam searches [36]. This comes with the *diverse subgroup set discovery* problem [46]: The extracted collection of k subgroups shall be of high quality and the less redundant as possible. Indeed, it often happens -if not always- that the best patterns differ very slightly on the objects they cover (or their description in a dual way). The redundancy makes that very few local optima are discovered and the subgroup set is not diversified. The solution proposed is to control each level of the beam search, which can be seen as a set of parallel hill-climbing searches[46,17]: The subgroup search space is explored level-wise and each level is restricted to a set of diversified high quality patterns. The diversification is done as follows. Subgroups are sorted according to their quality: The best one is picked and all the next patterns that are too similar (according to similarity of their covers and a threshold) are removed. The first of the next patterns that is not similar is kept, and the process is reiterated.

This problem of diversity with heuristic approaches has been thus well studied in SD/EMM, trying to make the best trade off between exploitation ("hill-climbing" nature of the beam-search) and exploration (diversity forced in the beam at each level). The problem we tackle in this article concerns diversity on the target space, which has not been studied to the best of our knowledge. Recall that we propose to enhance EMM by considering all target subspaces: It may happen that a subspace contains many subgroups of high quality and diversified in that subspace (the subgroups are very different). We need thus here as well to exploit promising subspaces while still exploring the others in order to cover many label combinations: This is what we call the *diversity in the target space*. To ensure the possibility of this diversity, the search space of subgroups should include the label sets. Our second contribution shows that ensuring the possibility of this diversity forces to reconsider the subgroup search space: Bisets $(subgroup, label\_set)$ are explored and should be wisely expanded during the search. We experiment this with the standard beam-search, but also with a very recent pattern sampling technique based on Monte Carlo Tree Search [6].

Finally, as our goal is to find rules where the consequent is a label set, we also need to reconsider the subgroup quality measure. Generally, one is

interested in rules with a high support and a few errors (maximizing the precision). The WRAcc is perfect for that as it expresses the difference between the precision of the subgroup w.r.t. the objects in the rest of the dataset, and is weighted with the subgroup support. However, in many applications, the label distribution is highly skewed. This is the case of our application in olfaction as it can be observed on Figure 4: Some label subsets are over-represented, others are under-represented. Even though, the neuroscientists are also interested in subgroups that involve such under-represented label subsets. Another solution would be to consider both the precision and recall, thus the F_1 measure. However again, this will favor small groups for which it is easier to find descriptions that cover well the label subset. In the best settings, one would favor precision for highly represented label subsets, and both precision and recall for under-represented subsets. We thus propose the F_β measure adapted for highly skewed label distributions, our third contribution. Actually, the F_β measure generalizes the F_1 measure in a way that β expresses a trade-off between the precision and the recall: We dynamically adjust it during the search, given the target subspace (label subset) currently enumerated. This directly impacts a better target diversity in the result set.

To summarize, our main contributions are manifold:

- We properly formalize a novel EMM instance to fully take into account multi-label data and all *target subspaces*.
- This comes with a more complex search space that needs even more to be handled with heuristic searches: We show how to adapt the subgroup search space so that heuristic search such as beam search and Monte Carlo tree search can be applied efficiently.
- We introduce several pattern quality measures that are able to take into account the skewness of the label distribution and dynamically consider the label distributions during the search to favor diversity on targets.
- We deeply experiment with these approaches on several benchmark multi-label datasets and an olfaction dataset. Our main result shows that MCTS with a relative F_β measure gives the best results in terms of computation times, quality and diversity both on the description and target spaces.
- We thoroughly demonstrate the actionability of the discovered subgroups for neuroscientists and chemists.

The rest of the paper is organized as follows. The next section recalls the basics of SD/EMM. Section 3 formalizes EMM with different target spaces for multi-label rule discovery. Section 4 introduces the subgroup search space and the several quality measures that favor target diversity. Section 5 develops the different heuristic algorithms. Section 6 covers the related work. Section 7 presents our experiments assessing the validity of our approach while Section 8 shows it actionability in neuroscience before to conclude.

2 Exceptional Model Mining and Diverse Subgroup Set Discovery

Definition 1 (Multi-label dataset) Let $\mathcal{O}$, $\mathcal{A}$ and $\mathcal{C}$ be respectively a set of objects, a set of attributes (either nominal or numerical), and a set of class labels. Each object is associated to a subset of class labels among $\mathcal{C}$ by the function $class : \mathcal{O} \mapsto 2^{\mathcal{C}}$ that maps the target labels to each object. We denote a multi-label dataset as $\mathcal{D}(\mathcal{O}, \mathcal{A}, \mathcal{C}, class)$.

Table 1 A toy dataset.

ID	a	b	c	$class(\cdot)$
1	150	21	11	$\{l_1,\ l_3\}$
2	128	29	9	$\{l_2\}$
3	136	24	10	$\{l_2,\ l_3\}$
4	152	23	11	$\{l_3\}$
5	151	27	12	$\{l_1,\ l_2\}$
6	142	27	10	$\{l_1,\ l_2\}$

Definition 2 (Subgroup) The description of a subgroup is given by $d = \langle f_1, \ldots, f_{|\mathcal{A}|} \rangle$ where each f_i is a restriction on the value domain of the attribute $a_i \in \mathcal{A}$. A restriction is either a subset of a nominal attribute domain, or an interval contained in the domain of a numerical attribute. The set of objects covering the description d is called the support of the subgroup, denoted $supp(d) \subseteq \mathcal{O}$. For simplicity, a subgroup is either given by its intent, i.e., its description, or by its extent, i.e., its support.

The aim of EMM is to find subgroups whose model over the class labels is significantly different from the model induced by the entire set of objects $\mathcal{O}$. Many models have been proposed in the literature [17]: The probability distribution, a Bayesian Network, a clustering or a classification model, etc., over the class labels. Quality measure have been introduced to compare the similarity between the two models (i.e., those of the subgroup and those of the entire dataset), the better the quality measure, the less the similarity, the more interesting the subgroup. Figure 1 displays the process of EMM. On the left, the red circles are the objects of the dataset. The green area is the support of a subgroup. The red curve on the middle is the model induced by the entire dataset. The green curve is the model induced by the subgroup. These two models are compared thanks to the quality measure.

Example 1 Consider the dataset in Table 1 with objects $\mathcal{O} = \{1, 2, 3, 4, 5, 6\}$, attributes $\mathcal{A} = \{a, b, c\}$ and class labels $\mathcal{C} = \{l_1, l_2, l_3\}$. Each object is labeled with a subset of class labels from $\mathcal{C}$. Considering the probability distribution model class over the class labels, the aim is thus to find subgroups whose distribution over all the class labels is significantly different from those of the entire set of objects. The distribution of the entire dataset for each class label l_1, l_2 and l_3 is respectively $p_0^{l_1} = \frac{|\{o \in \mathcal{O} | l_1 \in class(o)\}|}{|\mathcal{O}|} = \frac{3}{6} = 0.5$, $p_0^{l_2} = 0.67$ and

$p_0^{l_3} = 0.5$. Let us consider the subgroup s with description $d = \langle [128 \leq a \leq 151], [23 \leq b \leq 29] \rangle$. Note that for readability, we omit restrictions satisfied by all objects, e.g. $[9 \leq c \leq 12]$, and thus we denote that $supp(\langle\rangle) = \mathcal{O}$. The support of d is $supp(d) = \{2, 3, 5, 6\}$. The model induced by this subgroup for each class label is $p_d^{l_1} = \frac{|\{o \in supp(d) | l_1 \in class(o)\}|}{|supp(d)|} = \frac{2}{4} = 0.5$, $p_d^{l_2} = 1$ and $p_d^{l_3} = 0.25$. To compare this two models, we choose to use the mean of the WRAcc measures for each label:

$$\varphi(s) = \frac{4}{6} \times \left(\frac{(p_d^{l_1} - p_0^{l_1}) + (p_d^{l_2} - p_0^{l_2}) + (p_d^{l_3} - p_0^{l_3})}{3} \right) = 0.03$$

Since the search space of subgroups is too large, it is now widely accepted that an exhaustive search of subgroups, even with efficient pruning techniques, is not tractable for EMM. Heuristic methods are employed, such as beam search. However, it comes with the issue of diversity of the subgroup set that is extracted. The question is: How to ensure a high diversity in the result set of a heuristic exploration of the search space. This point has been studied as the Diverse Subgroup Set Discovery [46]. The aim is to extract a diverse subgroup set that is as small as possible to be easily interpretable by the experts. For that, a similarity measure (e.g., Jaccard coefficient between the support of two subgroups)is used to avoid the extraction of redundant subgroups within the result set.

Problem 1 (Diverse Subgroup Set Discovery [46]) Given a dataset $\mathcal{D}(\mathcal{O}, \mathcal{A}, \mathcal{C}, class)$, a quality measure φ , a minimum support threshold $minSupp$, an integer k, and a similarity measure sim, DSSD aims at extracting the diverse set of top-k best frequent subgroups w.r.t. the quality measure φ in which there is no similar subgroups w.r.t. sim.

3 Subgroup Set Discovery with Diversity on Target Subspaces

By definition, with the EMM framework, the model induced by the subgroup (and those induced by the entire set of objects) is always built over *all* the class labels. Thus, each subset of objects (or subgroup) derives a unique model. However, a subgroup can be deemed interesting only for the model induced over a subset of class labels because it derives a model completely different from those of the entire set of objects just for this subset of class labels. But, this subset of class labels is unknown *a priori*. Thus, it is required to explore the label set space, called the *target subspaces*. This is one of our contribution: We design a new EMM instance to strive subgroups whose model induced over an unknown subset of class labels $L \subseteq \mathcal{C}$ is different from the model induced by the entire dataset over the same subset of class labels L. The process of this new EMM instance is given in Figure 2. The change relies on the construction of the model. There is no longer *one* but several models derived from the subgroup: one for each target subspace. Thus, we need to refine Definition 2 about subgroups for this new EMM instance.

Definition 3 (Subgroup in a target subspace) Given $\mathcal{D}(\mathcal{O}, \mathcal{A}, \mathcal{C}, class)$, a subgroup, denoted $s = (d, L)$, is given by its description d and the subset of class labels $L \subseteq \mathcal{C}$ over which the model is built.

We will use the term *subgroup* in the rest of the article though, as it will be always implied that a subgroup is considered in a target subspace. Following the studies in *Diverse Subgroup Set Discovery*, this new EMM instance that deals with multi-label data should also exhibit a great diversity in the target subspace. Although it may happen that a subspace contains many subgroups of high quality and diversified in that subspace (the subgroups are very different), we need as well to exploit promising target subspaces while still exploring the others in order to cover many label combinations: This is what we call the diversity in the target space.

Problem 2 (DSSD with diversity on target subspaces) Given a dataset $\mathcal{D}(\mathcal{O}, \mathcal{A}, \mathcal{C}, class)$, a quality measure φ, a minimum support threshold $minSupp$, an integer k, and a similarity measure sim, DSSD aims at extracting the diverse set of top-k best frequent subgroups w.r.t. the quality measure φ in which there is no similar subgroups w.r.t. sim *and that covers as many target subspaces as possible.*

We will still refer to this problem as DSSD, as it implies in the rest of the article that diversity is considered both on the description and target subspace.

4 Quality Measures Considering the Target Subspaces

Evaluating a subgroup $s = (d, L)$ is performed thanks to a quality measure φ that computes the difference between the model induced by $supp(d)$ over the target subspace L and those induced by $\mathcal{O}$ over L. Complete surveys help understanding how to choose the right measure [19]. One of the most widely used quality measure for multi-label data is the Weighted Kullback Leibler divergence (WKL) [46]. The WKL of a subgroup $s = (d, L)$ is given by:

$$WKL(d, L) = \frac{|supp(d)|}{|\mathcal{O}|} \sum_{l \in L} (p_d^l \log_2 \frac{p_d^l}{p_0^l})$$

This measure aims at assessing the deviation between two distributions, i.e., it does not consider only the presence of a label, but also takes into the under-representation of a label for a subgroup. Moreover, WKL assumes that the labels are independent: It does not consider the co-occurrences of the labels. This is a strong assumption not verified in most of the data. Besides, note that WKL is maximized with $L = \mathcal{C}$ since it is the sum over the labels in L of positive terms. Thus, WKL can not be used in the settings of Problem 2. In this section, we study different existing measures that can be used for Problem 2 and detail their limits. Finally, we define a novel quality measure and its variants.

4.1 $WRAcc$ to evaluate the subgroups

The Weighted Relative Accuracy ($WRAcc$) is a well-known quality measure in EMM. Indeed, it allows to compare the proportion of a subset of labels in a subgroup with the proportion of this subset of labels in the entire dataset. For a subgroup $s = (d, L)$, it is given by:

$$WRAcc(d, L) = \frac{|supp(d)|}{|\mathcal{O}|} \times \left(p_d^L - p_0^L\right)$$

where $p_d^L = \frac{|\{o \in supp(d) | class(o) \subseteq L\}|}{|supp(d)|}$ (resp. $p_d^L = \frac{|\{o \in \mathcal{O} | class(o) \subseteq L\}|}{|\mathcal{O}|}$) is the proportion of objects in the subgroup s (resp. in the entire dataset) that are associated to all the labels in L. In other words, $WRAcc$ is the difference between the precision of the rule $d \rightarrow L$ and those of rule $\langle\rangle \rightarrow L$: the model of a subgroup s is given by the precision of the subset of labels L in $supp(s)$. This difference is weighted by the relative size of the subgroup to avoid the extraction of small subgroups. Note that, we can also consider the Relative Accuracy measure ($RAcc$) that does not weight the difference:

$$RAcc(d, L) = p_d^L - p_0^L$$

However, in our case study, we are interested in both the precision and the recall of the subgroup. $WRAcc$ or $RAcc$ only foster on the precision of the subgroup. We can notice that the weighted factor allows *in part* to take into account the recall by fostering on larger subgroups.

Example 2 Let us consider the dataset given in Table 1. For the description $d = \langle[128 \leq a \leq 151], [23 \leq b \leq 29]\rangle$ we can induce 7 different models, one for each subset of $\mathcal{C}$, namely $\{l_1\}$, $\{l_2\}$, $\{l_3\}$, $\{l_1,\ l_2\}$, $\{l_1,\ l_3\}$, $\{l_2,\ l_3\}$ and $\{l_1,\ l_2,\ l_3\}$. With the WRAcc measure: $\varphi(d, \{l_1\}) = \frac{4}{6} \times (\frac{2}{4} - \frac{3}{6}) = 0$, $\varphi(d, \{l_2\}) = 0.22$, $\varphi(d, \{l_3\}) = -0.17$, $\varphi(d, \{l_1, l_2\}) = 0.11$, $\varphi(d, \{l_1, l_3\}) = -0.11$, $\varphi(d, \{l_2, l_3\}) = 0.06$ and $\varphi(d, \{l_1, l_2, l_3\}) = 0$. Thus the best model induced by the description d is obtained for the subset of labels $\{l_2\}$.

4.2 F_1 score to take into account both precision and recall

The $F1$ score considers both precision $\left(P(d, L) = \frac{|supp(d) \cap supp(L)|}{|supp(d)|}\right)$ and recall $\left(R(d, L) = \frac{|supp(d) \cap supp(L)|}{|supp(L)|}\right)$ of a subgroup $s = (d, L)$. The Relative F_1 (RF_1) is given by:

$$RF_1(d, L) = F_1(d, L) - F_1(\langle\rangle, L)$$

where $F_1(d, L) = (2) \times \frac{P(d,L) \times R(d,L)}{(P(d,L)) + R(d,L)}$. Indeed, objects are described by both attributes and class labels, so the $F1$ score quantifies both the precision and the recall of the support of the description w.r.t. the support of the class labels.

Moreover, we can also consider the Weighted Relative F_1 (WRF_1), that uses the relative support size of the subgroup to weight RF_1:

$$WRF_1(d, L) = \frac{|supp(d)|}{|\mathcal{O}|} \times RF_1(d, L)$$

However, in most of the datasets, the distribution of class label is unbalanced. Some labels are associated to many objects in the dataset, and others are rarely used to label the data. Taking into account this setting is essential in this new EMM instance because the quality measure of the subgroups related to a frequent subset of class labels can be biased. F_1 does not equally evaluate subgroups related to over-represented subset of labels and subgroups related to not over-represented labels. Experimentally, we demonstrate that RF_1 is not effective to discover subgroups related to over-represented labels.

Example 3 Again, let us consider the toy dataset in Table 1 and the subgroup with description $d = \langle [128 \leq a \leq 151], [23 \leq b \leq 29] \rangle$. With $L = \{l_2\}$, the model induced by d is $F_1(d, \{l_2\}) = 1$. The model induced by the entire dataset is $F_1(\langle\rangle, \{l_2\}) = 0.8$. Thus, $\varphi(d, \{l_2\}) = RF_1(d, \{l_2\}) = F_1(d, \{l_2\}) - F_1(\langle\rangle 0, \{l_2\}) = 0.2$. With $L = \{l_1, l_2\}$, the models are $F_1(d, \{l_1, l_2\}) = 0.66$ and $F_1(\langle\rangle, \{l_1, l_2\}) = 0.5$, and thus $RF_1(d, \{l_1, l_2\}) = 0.16$. Using WRF_1 as quality measure, the result is $WRF_1(d, \{l_2\}) = 4/6 \times RF_1(d, \{l_2\}) = 0.13$ and $WRF_1(d, \{l_1, l_2\}) = 4/6 \times RF_1(d, \{l_1, l_2\}) = 0.11$

4.3 An adaptive F_β-score for skewed label distributions

The *WRAcc* focuses, by definition, too importantly on the precision of labels and experimentally returns subgroups in very populated target subspaces. The F_1 score makes a trade-off between precision and recall but we demonstrate experimentally that this measure promote subgroups covering few objects with low frequency label combinations. The trade-off lies with the so called F_β score which adds a parameter β that allows to tune the importance of the recall.

We propose β to be function of the support $\beta(|supp(L)|)$ of the considered target subspace L , so that the trade-off is automatically adapted during the search. The greater $|supp(L)|$, the closer to zero β is: The precision is fostered in F_β for *over-represented* labels. Conversely, the lower $|supp(L)|$, the closer to one β is: F_β is equivalent to the traditional F_1 score (the harmonic mean of precision and recall) for *non over-represented* labels. Formally, given two positive real numbers x_β and l_β, we define the Relative F_β score (RF_β) as follows (see also Figure 5):

$$RF_\beta(d, L) = F_\beta(d, L) - F_\beta(\langle\rangle, L)$$

where $F_\beta(d, L)$ is defined as follows:

$$F_\beta(d, L) = (1 + \beta(|supp(L)|)^2) \times \frac{P(d, L) \times R(d, L)}{(\beta(|supp(L)|)^2 \times P(d, L)) + R(d, L)}$$

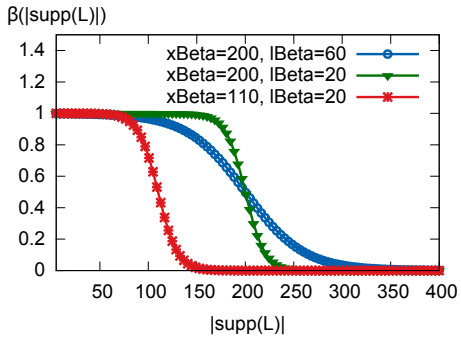


Fig. 5 The curves of $\beta(|supp(L)|)$.

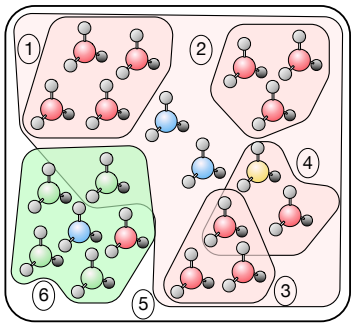


Fig. 6 Necessity of an adaptive measure.

where $\beta(|supp(L)|)$ is:

$$\beta(|supp(L)|) = 0.5 \times \left(1 + \tanh\left(\frac{x_\beta - |supp(L)|}{l_\beta}\right)\right)$$

Intuitively, for over-represented labels in the data, since it is difficult to find rules with high recall and precision, the experts prefer to foster the precision instead of the recall. They prefer extracting several small subgroups with a high precision than a huge subgroup (d, L) with plenty of non-L objects. In Figure 6 the red molecules are over-represented in the dataset, but it is more interesting having the different subgroups 1, 2, 3 and 4 with high precision, rather than a single huge local subgroup (5) which precision is much lower. For molecules that are not over-represented, the measure considers both precision and recall: e.g., subgroup 6 is possible for the green molecules.

Similarly to WRF_1 we can define the Weighted Relative F_β score as follows:

$$WRF_\beta(d, L) = \frac{|supp(d)|}{|\mathcal{O}|} \times RF_\beta(d, L)$$

Example 4 Let us take the example of the toy dataset in Table 1, with the description $d = \langle [128 \leq a \leq 151], [23 \leq b \leq 29] \rangle$ again. We consider the F_β measure to evaluate a model and RF_β to evaluate the subgroup. Since $|supp(l_1)| = 3$, $|supp(l_2)| = 4$ and $|supp(l_3)| = 3$, we fix $x_\beta = 3.3$ and $l_\beta = 0.5$. The β value for the subset of labels $\{l_2\}$ is 0.06 since in this setting, l_2 is over-represented within the dataset. The model induced by d on $L = \{l_2\}$ is $F_\beta(d, \{l_2\}) = 1$. The model induced by the entire dataset is $F_\beta(\langle\rangle, \{l_2\}) = 0.67$. Thus, $\varphi(d, \{l_2\}) = RF_\beta(d, \{l_2\}) = F_\beta(d, \{l_2\}) - F_\beta(\langle\rangle, \{l_2\}) = 0.33$. With WRF_β, we have $WRF_\beta(d, \{l_2\}) = \frac{4}{6} \times RF_\beta(d, \{l_2\}) = 0.22$ For the subset of class labels $\{l_1, l_2\}$, $\beta = 0.99$. The models are $F_\beta(d, \{l_1, l_2\}) = 0.66$ and $F_\beta(\langle\rangle, \{l_1, l_2\}) = 0.5$, and thus $\varphi(d, \{l_1, l_2\}) = 0.16$. Using WRF_β, the quality measure of s is $WRF_\beta(d, \{l_1, l_2\}) = \frac{4}{6} \times RF_\beta(d, \{l_1, l_2\}) = 0.11$. Note that for $L = \{l_1, l_2\}$, RF_β and WRF_β are equivalent to RF_1 and WRF_1 since L is not over-represented in the data.

5 Search space explorations

We present the search space of subgroups that needs to be traversed in all target subspaces. We briefly detail an algorithm for exhaustive search as a baseline for some of our experiments. We then present the two heuristic search techniques that we employ in our experiments.

5.1 Search space and exhaustive search

In standard SD and EMM, the search space of subgroups is given by the lattice of all possible descriptions $(D, \sqsubseteq)$ where $d_1 \sqsubseteq d_2$ means that subgroup d_1 is more general than d_2, or equivalently $supp(d_2) \subseteq supp(d_1)$. This lattice can be explored either in a depth-first (DFS) or in a breadth-first (BFS) search manner. During the traversal, the quality measure is computed for each subgroup. In the end, a redundancy filter is applied to output the top-k diverse subgroups. This filter is explained in the next subsection.

In our case, we wish to evaluate a description on each of the target subspaces. We need to consider the following search space: $D \times 2^{\mathcal{C}}$, where $\mathcal{C}$ is the set of labels. Hence, a subgroup is always considered in a target subspace in which the quality measure can be computed. Thus, only slight modifications in existing algorithms are required. We override the specialization/generalization relation $\sqsubseteq$ as follows: $(d_1, L_1) \sqsubseteq (d_2, L_2) \iff supp(d_2) \subseteq supp(d_1) \wedge L_2 \subseteq L_1$

For that, we adapt the algorithm *CloseByOne* [31] from the Formal Concept Analysis [20] that can handle easily both nominal and numerical attributes [24]. Without entering into the details, it avoids to generate subgroups having exactly the same support and truly operates an exhaustive search. Indeed, we could have adapted the most efficient subgroup discovery algorithm, SDMap* [2], but it is not purely exhaustive as it operates greedy cutting of numerical attributes. Moreover, we focus on heuristic search and already shown that MCTS performs better than SDMap* for large search space in a previous work [6] (considering the WRAcc, only on basic subgroup discovery).

5.2 Heuristic search with Beam-search

Beam search [36] is the most popular heuristic technique in SD/EMM. It has been originally adapted to consider the *diverse subgroup set discovery* problem by Leeuwen and Knobbe [46] as follows. The subgroup search space is explored level-wise (BFS) and each level is restricted to a set of diversified high quality patterns. The diversification is done as follows. Subgroups are sorted according to their quality: The best is picked and all the next patterns that are too similar (bounded Jaccard coefficient between their support) are removed. The first of the next patterns that is not similar is kept, and the process is reiterated.

Adapting beam search considering the search space $D \times 2^{\mathcal{C}}$ with diversity also on the target subspaces is done as follows. It starts from the most general subgroup. Next levels are generated by specializing subgroups either by

restricting an attribute or by extending the subset of class labels with a new label it has also to characterize as long as the quality measure is improved. There are at most $|\mathcal{C}| + \sum_{a_i \in \mathcal{A}} |a_i|(|a_i|+1)/2$ possibilities to specialize each subgroup: We can proceed up to $|\mathcal{C}|$ extensions of the subset of labels to characterize L and $|\mathcal{A}|$ extensions of the description for which we can build up $|a_i|(|a_i|+1)/2$ possible intervals for numeric attributes. We choose among those only a constant number of candidates to continue the exploration (the beam width: The $beamWidth$ best subgroups w.r.t. the quality measure). Removing the redundancy is done as in the classical case, but each subspace is considered separately in order to avoid to favor few excellent target subspaces: the subgroups are split into groups according to their target subspace, and the diversity filter is operated on each of them.

5.3 Sampling patterns with Monte Carlo Tree Search

We propose to sample the subgroup search space $X \times 2^{\mathcal{C}}$ relying on Monte Carlo Tree Search [8]. Indeed, in a previous work [6], we showed that MCTS is able to handle very large search space and outperforms beam search in terms of quality and diversity of the pattern set when considering basic subgroup discovery and the WRAcc. Moreover, MCTS is a budget based approach: The more time and memory allocated, the better the result.

Without entering into the details, MCTS iteratively draws a random subgroup description d_n, following a path $d_0 \sqsubset d_1 \sqsubset d_2 \sqsubset ... \sqsubset d_n$. The best pattern quality measure found on the path is returned as a reward. d_0 is stored in a memory (the Monte Carlo tree) and the reward backpropagated in the tree: Each node stores the number of times it was visited and the average quality measure obtained so far. The tree will drive the search for the next iterations (thanks to the upper confidence bound, a formula that expresses a trade-off between exploration and exploitation of the search space [29]). In other terms, a new d_0 will be generated according to this tradeoff.

More specifically, MCTS operates a fixed number of iterations where: (i) it *selects* the most urgent node d_{-1} in the tree according to the UCB; (ii) it *expands* this node by randomly selecting one of the direct description specialization $d_{-1} \sqsubset d_0$, (iii) it *simulates* a random path $d_0 \sqsubset d_1 \sqsubset d_2 \sqsubset ... \sqsubset d_n$, and (iv) the best reward φ found on the path is used to *update* the tree. This is illustrated in Figure 7. Each of these four steps can be achieved with many strategies. We invite the reader interested by this approach to consult our research report [6]: We use the best settings found for the WRAcc in basic subgroup discovery settings.

6 Related work

Subgroup Discovery (SD) was first introduced in the middle of the 90's and is related to supervised learning [28,48]. The input data is a population of

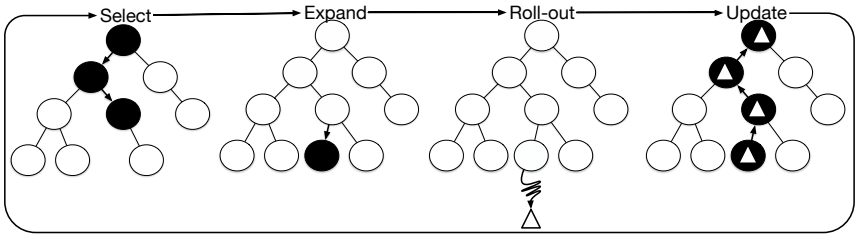


Fig. 7 Monte Carlo Tree Search Principle.

individuals (e.g., objects, customers, transactions) that embeds a set of descriptors and a target variable (a label or a numeric value). The aim of SD is to extract subgroups of individuals (described by a rule involving descriptors) for which the distribution on the target variable is statistically different from the whole (or, for some authors, the rest of the) population. Although these settings are related to those of a supervised learning task, SD is a descriptive task. Since that, two other notions have been formalized gathering the same settings: Contrast Set mining [3] and emerging patterns [14]. The first one aims to obtain high differences of support between the values of the target variable. The latter extracts patterns with different frequencies in two classes (e.g., the positive and the negative classes). However, these both methods and SD are similar and differ partially from the quality measure they use [40]. All these techniques apply kinds of supervised learning to descriptive tasks.

Exceptional Model Mining (EMM) was first introduced by Leman et al. [33]. Recently, Duivesteijn et al. have proposed a deep survey on EMM [17]. EMM can be seen as a generalization of traditional SD. Indeed, EMM enables to deal with more complex target concepts. Initially, SD aims at finding subgroups of objects for which the distribution over one class label deviates substantially from the distribution of the entire population of objects. In EMM, there is no longer one target variable but several ones: The variables are split into two sets, the descriptive variables and the target labels. The goal of EMM is to extract subgroups of objects for which the model induced over all the target labels substantially deviates from the model induced by the whole population of objects.

EMM is based on a class model. Lots of model classes have been identified by the community (linear regression, contingency tables, Bayesian networks, ...). The choice of the model class depends on both the target attributes and the purpose of the application. The main stream of works about EMM consists of finding new model classes to deal with specific datasets and objectives [17]. The simplest models deal with the correlation (or association) between two target attributes: The aim is to find subgroups for which the correlation is significantly different from those induced on the entire dataset. One can also be interested in the difference of the linear regression of one numerical target attribute in the subgroups, by comparing for instance the slope of these models. More complex model can also be suitable for specific cases:

e.g., the models based on Bayesian networks. In this model, it is assumed that there are multiple nominal target attributes. A subgroup is deemed interesting if the conditional dependencies relations between the target attributes are significantly different in the subgroup from those of the entire dataset [18]. Recently, a new model has been proposed to handle ranking data. The Rank Correlation Model Class has been introduced to be less sensitive to the outliers in the target attributes [15]. This class model no longer uses the values taken by the target attributes but their rank. As a last example, the exceptional aspect of a subgroup can also be computed on a graph induced by a subgroup compared to the graph modelling the whole [25].

Once the model is chosen, it is required to select the quality measure that compares two instances of the model class. The quality measure is the heart of the method in SD or EMM. It depends on both the model and the purpose of the application. There exists a large panel of quality measures in the literature [18,16,30,17]. Usually, when the application is specific, it requires to design a new quality measure that enables to encode the needs of the experts.

Several algorithms have been proposed to explore the search space, i.e. the subgroup space structured as a lattice of the descriptions. The first algorithms performed an exhaustive exploration of the search space [28,48,23]. More recent works led to efficient exhaustive exploration: *SD-MAP** [2] (and its extension to EMM, called *GP-Growth* [35]) employs the FP-Growth method [22], *Merge-SD* is based on bounds to prune [21], and another approach based on optimistic estimates on different quality measures [34]. However, the exhaustive search is not suitable for large datasets, and the use of heuristics is required. For that, most of the SD/EMM algorithms employ a beam search strategy [36, 46,38]. Another trend of heuristic search for EMM is about sampling methods. A recent work employs Controlled Direct Pattern Sampling (CDPS) to search for subgroups in the EMM framework [39]. CDPS is a sampling method that enables to create random patterns thanks to a procedure based on a controlled distribution [5]. In a recent work [6], we showed that Monte Carlo Tree Search is a method that can sample subgroups very efficiently, relies on a exploration/exploitation trade-off, outperforms beam search in terms of diversity, and most importantly, it is agnostic of the description language and does require any assumption about the quality measure (in contrast with other pattern sampling techniques). Finally, genetic algorithms have been proposed to deal with SD/EMM [12,42,4,10]. These approaches aim at solving problems imitating the process of natural evolution but hardly scale to high dimensional data.

The redundancy is one of the main issue for SD/EMM. Indeed, once the subgroups have been extracted, the result set contains a lot of duplicates (similar descriptions and supports). To avoid the redundancy in the result set, many works of the state of the art proposed to filter out redundant subgroups. The first attempt is performed by van Leeuwen and Knobbe [45]. They proposed a similarity measure to filter out redundant subgroups based on the similarity of the descriptions and the supports of the subgroups. To go further, van Leeuwen and Ukkonen also proposed to used skylines methods to deal with redundant

subgroups [47]. In fact, considering the Pareto front ensures to avoid some of the redundant subgroup of the result set.

Finally, the closest work we found in the literature defines a list of 8 types of rules in single and multi-label data [37]. The subgroups we consider in our work are equivalent to their so called *Multi-label and sparse label-independent rules* but they focus on rules where only a single label can appear in the head of the rule and other label can appear in the body. They opt for a kind of binary relevance transformation (each label is learned separately).

7 Experiments

Experiments were carried out on an Intel(R) Core(TM) i7-7700HQ CPU 2.80 GHz machine with 8 GB RAM. All materials are available on `https://github.com/BelfodilAimene/dssd-fbeta` After introducing the different datasets, we will answer to the following questions:

- How to tune x_β and l_β parameters of the F_β measure?
- Is the MCTS approach also able to consider with success the (Weighted) Relative F_β, as it was shown only to performs well with the $WRAcc$ in our previous work [6]?
- What is the best measure to ensure target subspace diversity with MCTS?
- Is that measure also the best for beam search? It will thus advocate that the measure is not dependent of the heuristic.
- Is the (weighted) relative F_β ranking the subgroups differently? It will thus experimentally show that the other measures are not equivalent.
- How are precision and recall of the subgroup distributed w.r.t. target subspace frequency? We should observe that the precision-recall trade-off is respected in the result sets: Fostering on precision for over-represented label sets, and also on the recall for the others.

7.1 Data

We experiment our approach with a set of multi-label datasets from the well-known MULAN repository [44]. We also experiment with an *olfaction* dataset built by a neuroscientist (co-author of the present work) which is deeply explained in the next section where we assess the interest of our approach in a real world scenario. The Table 2 gives the properties of each dataset: The number of instances $|\mathcal{O}|$, the number of attributes $|\mathcal{A}|$, the domain of the attributes (nominal or numeric), the number of labels $|\mathcal{C}|$, average number of labels associated to an object (cardinality), its density (the cardinality divided by the number of labels) and the mean and median of the number of objects per label. Figure 8 finally gives the label distribution for each dataset.

Dataset	$\|\mathcal{O}\|$	$\|\mathcal{A}\|$	*Domain*	$\|\mathcal{C}\|$	*cardinality*	*density*	*distinct*	*mean*	*median*
CAL500	502	68	numeric	174	26	0.15	502	75.14	39
emotions	593	72	numeric	6	1.8	0.31	27	184.67	170
enron	1702	1001	nominal	53	3.3	0.06	753	108.49	26
genbase	662	1186	nominal	27	1.2	0.04	32	30.70	17
yeast	2417	103	numeric	14	4.2	0.30	198	731.5	659
Olfaction	1689	82	numeric	74	2.882	0.04	1069	65.80	28

Table 2 Datasets used for the experiments.

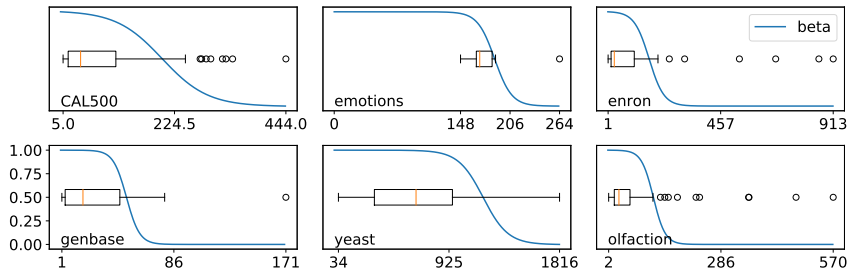


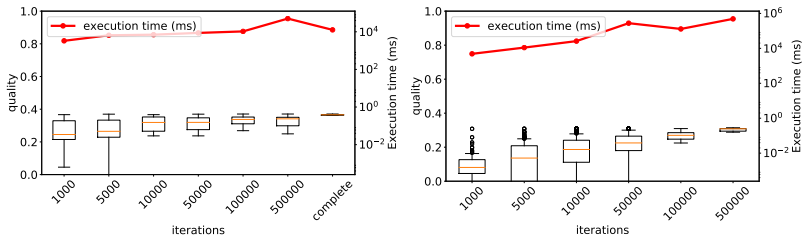
Fig. 8 Label distribution in the different datasets of Table 2.

7.2 How to choose the x_β and l_β parameters for F_β?

Using F_β requires to set the two parameters x_β and l_β. In order to choose these two parameters, one needs to answer to the following questions: At what *support size* a label subspace L is considered as over-represented? What is the speed of transfert from the normal representation state to the over-represented state? They are set thanks to the characteristics of the dataset. An automatic approach is to set x_β to the 85^{th} percentile value and l_β to the difference between the 85^{th} and 80^{th} percentiles. However, we advise to display the distribution of the frequency of the class labels (as shown in Figure 8) to correctly set up these parameters. The user must keep in mind that when $|supp(L)| = x_\beta + 2l_\beta$, $\beta \approx 0.02$ (L is considered thus over-represented and precision is fostered). Analogically, when $|supp(L)| = x_\beta - 2l_\beta$, $\beta \approx 0.98$ (L is considered normal and the F_β worth almost $F1$ score).

7.3 Is MCTS able to consider other measure than $WRAcc$?

Since the search space related to datasets we use are too large to employ an exhaustive search, we design two heuristic method to experiment with the DSSD on diverse target subspaces problem. In this subsection, we show that MCTS using RF_β allows to extract a diverse set of non redundant patterns if given enough computational budget (e.g., number of iterations). Note that,



(a) Reduced Olfaction (b) Yeast

Fig. 9 Evolution of the quality of the result set of MCTS varying the number of iterations.

in a previous work, we argued that MCTS enables to efficiently extract interesting subgroups with existing measures such as $WRAcc$ and F_1 [6]. Figure 9 displays the evolution of the quality of the result set with MCTS varying the number of iterations. We can note that, the more the iterations, the better the quality in the result set. Moreover, we use an exhaustive search in Olfaction to show that MCTS quickly converges to the quality of the result set obtained with the exhaustive search. However, since the search space of this dataset is too large, we randomly picked 2 attributes among the 82 attributes to make the exhaustive search tractable. Figure 9(a) shows that, given enough computational budget, MCTS converges to the quality of the result set obtained by the exhaustive search. From that, it is legitimate to employ MCTS with this new quality measure RF_β. Furthermore, on Figures 9(b)-(c), we can notice that the quality of the result set still increases with the number of iterations.

7.4 Which measure ensures the most diverse result?

We compare the diversity of the target subspaces in the result set. For that, we use MCTS with 100k iterations. We output the top-1000 subgroups when using the different quality measures $RAcc$, RF_1, RF_β, $WRAcc$, WRF_1 and WRF_β. Figures 10-11 display the result we obtain. We observe that, in general, $RAcc$ and $WRAcc$ covers few label subspaces compared to RF_1, WRF_1, RF_β and WRF_β. Indeed, in Olfaction, $RAcc$ covers twice less label subspaces than the other (see Figure 10(d)-(e)-(f)). However, on few datasets, the diversity on the target subspaces is almost the same, e.g., Figure 10(d)-(e)-(f) for Genbase.

Besides, contrary to RF_1, the measures $RAcc$ and RF_β are able to homogeneously evaluate the subset that are over-represented and the non-over-represented. Thanks to the adaptable F_β, RF_β can either support the precision for the over-represented labels, or both precision and recall for non-over-represented labels. For the non-over-represented labels $RAcc$ does not also foster on the recall of the subgroups. However, with RF_1, the over-represented subgroups are rarely output in the result set because their recall is low when the precision is high and vice versa.

Moreover, the experiments with the weighted version of these measures lead almost to the same result: WRF_β ensures a great diversity in the target subspaces and is able to characterize both over-represented labels and non-over-represented label. However, the weighted factor fosters more on over-represented labels since the relative support size is greater. Note that when the distribution of the labels is not imbalanced, such as in Emotions, the measures behaves in the same way, which is an expected behavior.

Finally, Figure 12 displays both the evolution of the diversity on the target space and the value of the quality measure of the subgroups in the result set varying the number k of output pattern. Figures 12(a)-(b)-(c) show that with *RAcc* or *WRAcc* there is a few diversity in the target subspace whereas with RF_1, WRF_1, RF_β and WRF_β the diversity is high. Moreover, Figures 12(d)-(e)-(f) display that with the weighted version of these measure, the quality in the result set decreases faster than with the relative version. Indeed, the weighted factor makes the quality measure foster on over-represented subgroups and thus avoid the extraction of non-over-represented even if the precision and the recall is high.

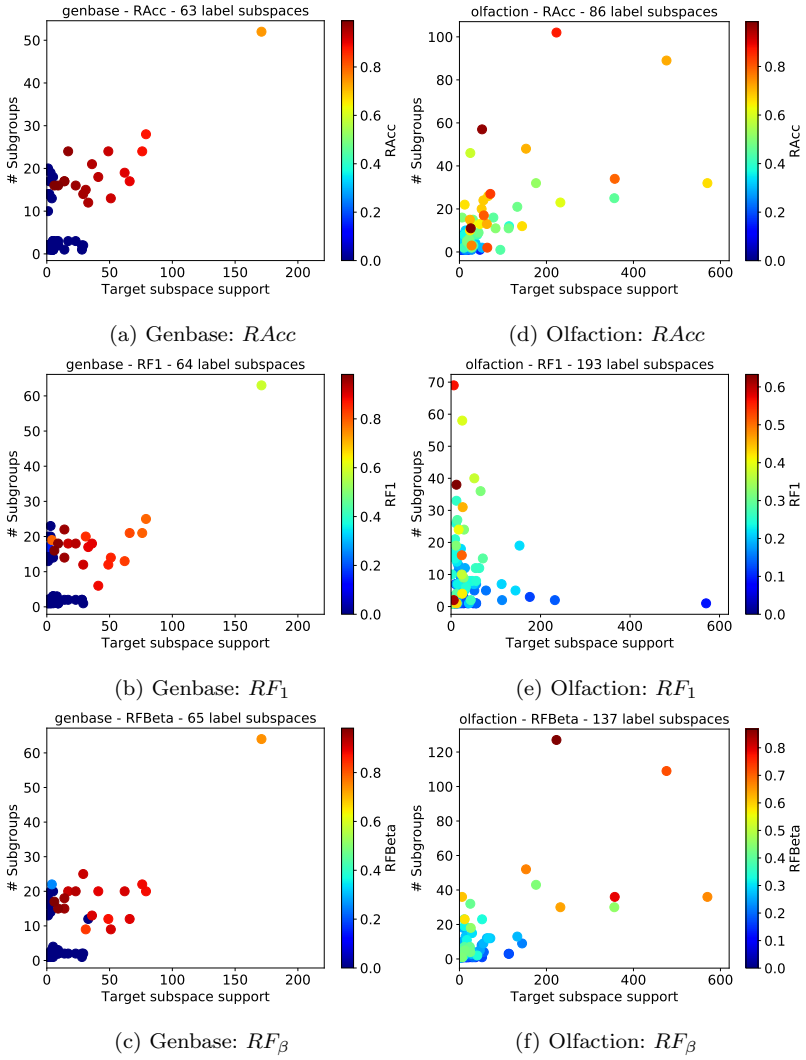
7.5 Does RF_β also ensure the best diversity?

The previous subsection exhibits that RF_1, WRF_1, RF_β and WRF_β increase the diversity on the target space in the result set. In some case, we showed that this diversity is also high with *RAcc* and *WRAcc*. However, in our previous work, we ensured that MCTS leads to a higher diversity [6]. Thus, in this subsection, we present some results obtained with a beam search. Figure 13 shows that for all the measures, except RF_β, there is a low diversity on the target space. Besides, except with RF_β, the beam search can not extract more than 300 diverse and non redundant subgroups. Clearly, RF_β provides the largest diversity in the result set.

7.6 Is the RF_β ranking the subgroups differently?

We experiment with the 6 different quality measures, to test if they are equivalent: i.e. they provide the same ranking of subgroups. Since it is not possible to use an exhaustive search, we can not directly compare the ranking of the subgroups (most subgroups in a result set are not included in another result set). For that, we define a optimistic similarity measure between two result sets based on Jaccard coefficients: we compute the mean of the maximum of the Jaccard coefficients between a couple of subgroups in the two different result set. Formally we compute $meanSim(S_1, S_2) = \frac{1}{|S_1|} \times \sum_{s_1 \in S_1} \max_{s_2 \in S_2} J(supp(s_1), supp(s_2))$ and then the similarity between two result sets S_1 and S_2 is:

$$maxSim(S_1, S_2) = \max\left(meanSim(S_1, S_2), meanSim(S_2, S_1)\right)$$



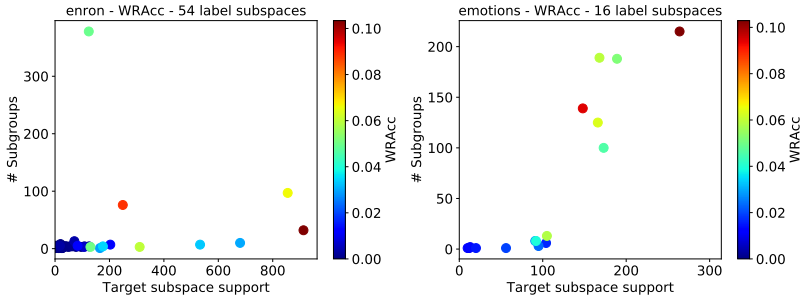
(a) Genbase: $RAcc$ (d) Olfaction: $RAcc$

(b) Genbase: RF_1 (e) Olfaction: RF_1

(c) Genbase: RF_β (f) Olfaction: RF_β

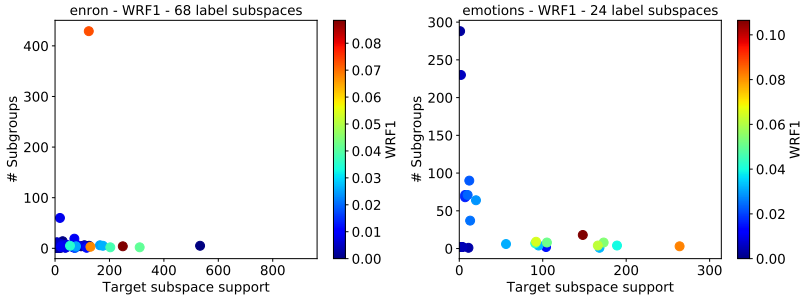
Fig. 10 The quality measure and the support of the target subspace of the subgroups within result set obtained with MCTS using the Relative measures. The color of the points is the value of the quality measure of the subgroup given by the heatmap.

Figure 14 displays the matrix of *maxSim* between the result sets obtained with the different quality measures in CAL500. Clearly, it shows that they do not share the same subgroups in their result set. Thus, the measures are not equivalent. The results are identical for the other datasets.



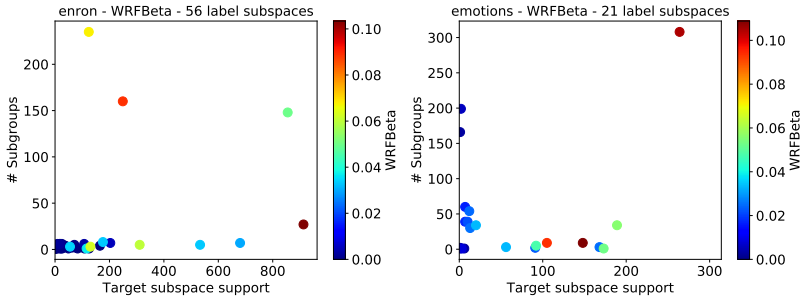
(a) Enron: $WRAcc$

(d) Emotions: $WRAcc$



(b) Enron: WRF_1

(e) Emotions: WRF_1



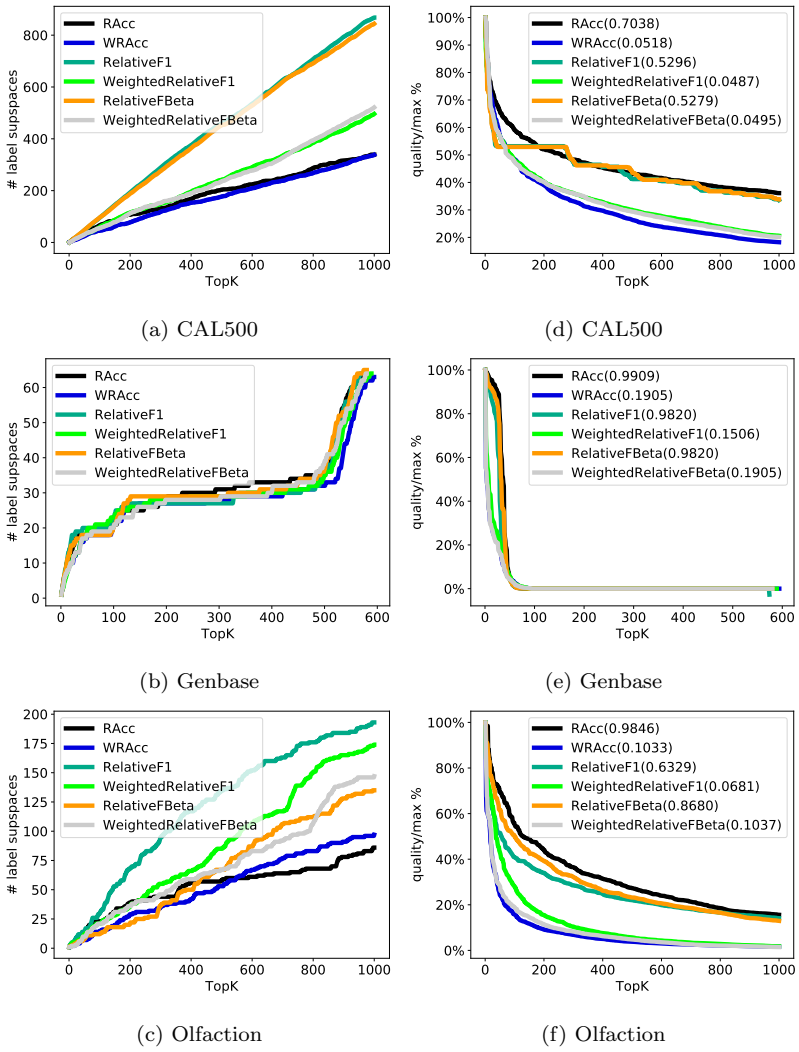
(c) Enron: WRF_β

(f) Emotions: WRF_β

Fig. 11 The quality measure and the support of the target subspace of the subgroups within result set obtained with MCTS using the Weighted Relatives measures. The color of the points is the value of the quality measure of the subgroup given by the heatmap.

7.7 Does F_β dynamically adapt to the label frequency?

We said that for the over-represented labels, we want to support the precision, and for other frequency of labels we take into account both the precision and the recall of the subgroup. To assess the behavior of the measure, we experiment with the measure we use to evaluate independently a model, namely Acc, F_1 and F_β. From that, we display the precision and the recall of the subgroups. Figure 15 shows two different views of the precision and the recall



(a) CAL500 (d) CAL500

(b) Genbase (e) Genbase

(c) Olfaction (f) Olfaction

Fig. 12 Comparison between the measures on the extracted top-K. Evolution of the quality measure in the top-K.

of the subgroup depending on the frequency of the subset of labels L they are related on the Olfaction dataset. Figures 15(a)-(b)-(c) show the precision (triangles) and the recall (crosses) of the subgroup in the result set. They also display the value of $\beta(|supp(L)|)$. Clearly, for F_β, if the frequency of L is too high in the dataset, only the precision of the subgroup is fostered and the recall are low. But, if the frequency of L is *not* too high, both recall and precision are fostered. However, with *Acc*, only the precision is fostered, the recall is always low whatever the frequency of L. Concerning F_1, we support both the

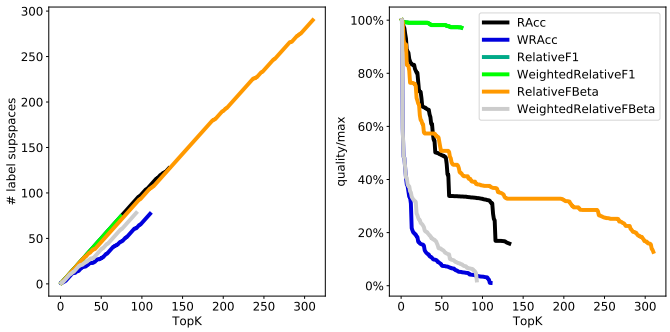


Fig. 13 Comparison with a beam search on the Olfaction dataset. .

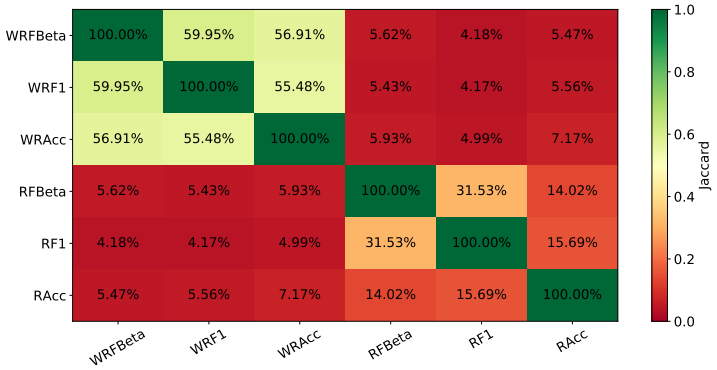


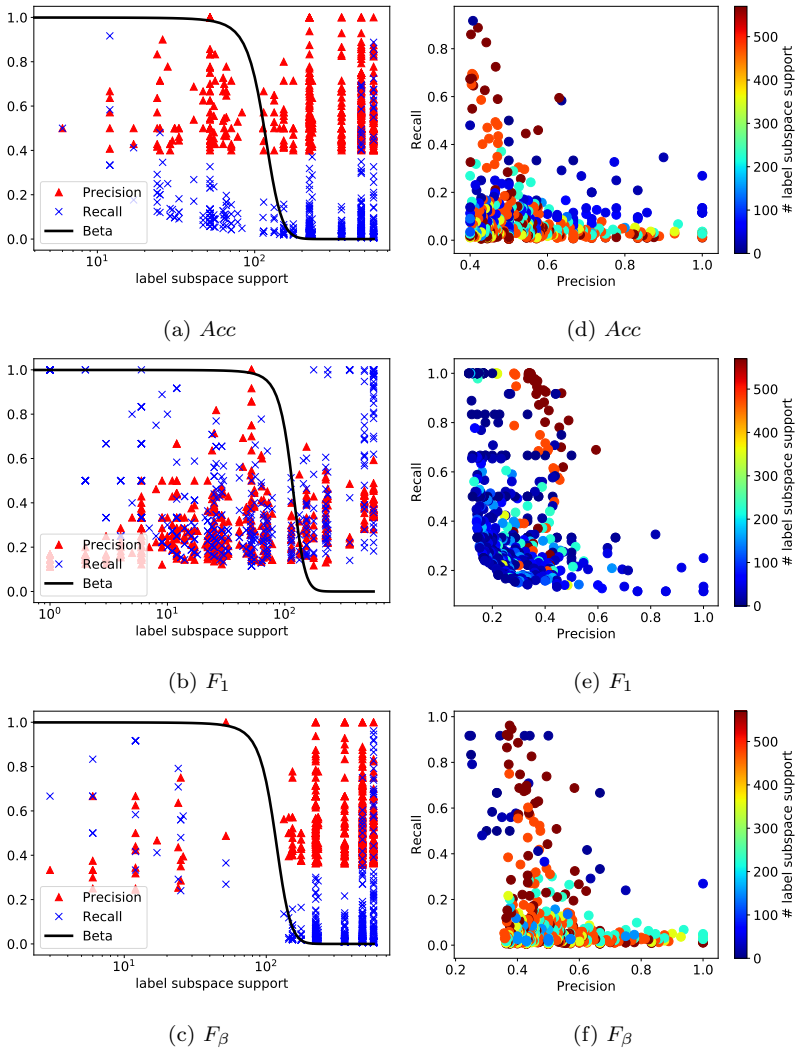
Fig. 14 The matrix of the similarity of the result set using different quality measures.

precision and the recall, but there are less subgroups in the result set that are related to labels for which the frequency is high because the F_1 score is low.

Figures 15(d)-(e)-(f) present another point of view to illustrate this behavior. The subgroups in the result set are plotted in the precision/recall space. Moreover a heatmap is used to show the frequency of the subset of labels L to which the subgroups are related: The points in red are subgroups related to over-represented labels, and the points in blue are subgroups related to under-represented labels. Clearly, Acc does not foster on the recall. Besides, F_1 extracts few subgroups related to over-represented labels (many blue points). Finally, F_β behaves as expected: (i) Both over-represented and under-represented labels are covered, i.e., the diversity in the target space is high ; and (ii)for over-represented labels the precision is fostered, and for other labels both recall and precision are taken into account.

7.8 Synthesis of the experiments

The parameters x_β and l_β for the F_β measure could be easily set thanks to the distribution of the labels frequency. In general, these results suggest that ex-



(a) *Acc* (d) *Acc*

(b) F_1 (e) F_1

(c) F_β (f) F_β

Fig. 15 Precision and recall of the models of the extracted subgroups on Olfaction.

haustive search are not tractable because the search spaces of the datasets are too large, and thus heuristic explorations are needed. MCTS is able to quickly find interesting subgroups with RF_β and WRF_β. Moreover, the results support the idea that RF_β provides a better diversity on the target subspace using either MCTS or a beam search. All these quality measures focus on different subgroups and thus there is no equivalence between each other. Finally, as expected the F_β measure enables to foster the precision for over-represented labels and support both recall and precision for other label subsets.

8 Application in neuroscience

8.1 Understanding the olfactory percept

Around the turn of the century, the idea that modern, civilized human beings might do without being affected by odorant chemicals became outdated: the hidden, inarticulate sense associated with their perception, hitherto considered superfluous to cognition, became a focus of study in its own right and thus the subject of new knowledge. It was acknowledged as an object of science by Nobel prizes (e.g., [9] awarded 2004 Nobel prize in Physiology or Medicine); but also society as a whole was becoming more hedonistic, and hence more attentive to the emotional effects of odors. Odors are present in our food, which is a source of both pleasure and social bonding; they also influence our relations with others in general and with our children in particular. The olfactory percept encoded in odorant chemicals contribute to our emotional balance and wellbeing.

While it is generally agreed that the physicochemical characteristics of odorants affect the olfactory percept, no simple and/or universal rule governing this Structure Odor Relationship (SOR) has yet been identified. Why does this odorant smell of roses and that one of lemon? Considering that the totality of the odorant message was encoded within the chemical structure, chemists have tried to identify relationships between chemical properties and odors. However, it is now quite well acknowledged that structure-odor relationships are not bijective. Very different chemicals trigger a typical "camphor" smell, while a single molecule, the so-called "cat-ketone" odorant, elicit two totally different smells as a function of its concentration [11]. At best, such SOR rules are obtained for a very tiny fraction of the chemical space, emphasizing that they must be decomposed into sub-rules associated with given molecular topologies [13]. A simple, universal and perfect rule does probably not exist, but instead, a combination of several sub-rules should be put forward to encompass the complexity of SOR.

In this paper, we propose a data science approach with a view to advance the state of the art in understanding the mechanisms of olfaction. Whereas it has been show recently that some odors can be predicted given the physico-chemical properties of the molecules [26], the most accurate methods generally never suggest a descriptive understanding of the classes, while fundamental neurosciences need descriptive hypotheses through exploratory data analysis, i.e., descriptions that partially explain SOR. This is where our EMM approach fully makes sense as molecules are associated to several odors, with a highly skewed distribution and a very large search space (82 numerical attributes with a large number of possible values).

8.2 An original olfaction dataset

One prominent methodological lock in the field of neuroscience concerns the absence of any large available database (>1000 molecules) combining odorant molecules described by two types of descriptors: perceptual ones such as olfactory qualities (scent experts defining a perceptual space of odors), and chemical attributes (chemical space). The dataset provided by the IBM challenge [26] is a clinical one: i.e., odorant molecules were not labeled by scent experts. To tackle this issue, the neuroscientists selected a list of $1,689$ odorants molecules described by 74 olfactory qualities in a standardized atlas [1]. They then described using *Dragon 6 software* (available on `talete.mi.it`) all of these molecules at the physicochemical levels (each odorant molecule was described by more than 4,000 physicochemical descriptors). As such, and to the best of our knowledge, the present database, created by neuroscientists, is one of the very few in the field that enable quantification and qualification of more than $1,500$ molecules at both, perceptual (neurosciences) and physicochemical (chemistry) levels. The distribution of the 74 olfactory qualities is illustrated in Figure 4. We use this dataset in the rest of this section after restricting the 4,000 attributes in different manners. The selection of 82 attributes used in the previous section has been made by a scientist expert in chemistry of odors (also co-author of this work). For three other selections, we filter out correlated attributes with the Pearson product-moment correlation coefficient: As a result, attributes with a correlation higher than 90% (resp 60% and 30%) were removed leaving only 615 (resp. 197 and 79) attributes. All experiments in the rest of this section were performed our web plateform called *Olfamining* [7].

8.3 Identification of relevant physicochemical attributes

We consider the experiment on Olfaction dataset when we use the RF_β score, $minSupp = 30$. A relevant information for neuroscientists and chemists concerns the physicochemical attributes that were identified in the descriptive rules. As showed in [27], the sum of atomic van der Waals volumes, denoted as Sv, is discriminant with regard to the hedonism of an odor, and especially the higher Sv, the more pleasant an odor. Moreover, the higher the rate of nitrogen atoms ($N\%$), the less pleasant an odor, consistent with the idea that amine groups ($-NH_2$) are associated with bad odors (such as cadaverine or putrescine). Based on this observation, we find subgroups related to either the *Floral* or *Fruity* quality that are characterized by a special range of values with regard to Sv and $N\%$. For example, $s_1 = \langle [3.0 \leq N\% \leq 33.3]\ [2.7 \leq O\% \leq 33.3]\ [0.0 \leq nR05 \leq 1.0]\ [nRCN = 0.0], \{Fruity, Grape\}\rangle$ and $s_2 = \langle [6.57 \leq Sv \leq 43.17]\ [1.0 \leq nN \leq 3.0]\ [1.0 \leq nHDon \leq 2.0], \{Floral\}\rangle$ are output subgroups. The quality measure of s_1 is 0.23 with a precision of 0.16 and a low recall of 0.5. For s_2, its quality measure is up to 0.37,its precision is 0.65 and its recall is 0.06. The first subgroup contains in its description the $N\%$ attribute associated to a very low percent-

age, and s_2 includes the Sv attributes with a range of values that corresponds to its higher values. In general, the quality $Musk$ is associated with large and heavy molecules: the molecular weight (MW) of these molecules is thus high. In the output subgroups, most of those associated to the musk quality include in their description the MW attribute with high values, or any other attribute that is positively correlated with MW, such as $SAtot$. For example, $s_3 = \langle [27.03 \leq MW \leq 297.3]\ [1.0 \leq nBM \leq 18.0]\ [nR03 = 0.0]\ [0.0 \leq nCp \leq 7.0]\ [5.0 \leq nCrs \leq 16.0]\ [nHDon = 0.0], \{Musk\}\rangle$ with a quality measure of 0.2 (precision: 0.27, recall: 0.25) is about molecules with it molecular weight between 27.03 and 297.3. Moreover, when the quality *Musk* is combined with the quality *Animal*, we still have a high molecular weight but there are other attributes with specific range of values: $s_4 = \langle [168.808 \leq SAtot \leq 464.918][6.0 \leq nCar \leq 16.0][3.261 \leq MLOGP \leq 4.593]\ [0.0 \leq nCsp2 \leq 7.0], \{Musk, Animal\}\rangle$. This latter topological attribute is consistent with the presence of double bonds (or so-called $sp2$ carbon atoms) within most musky chemical structure, that provides them with a certain hydrophilicity.

8.4 Providing relevant knowledge to solve a theoretical issue in the neuroscience of chemo-sensation

Another important information brought by these findings to experts lies in the fact the SOR issue should be viewed and explored through a "multiple description" approach rather than "one rule for one quality" approach (i.e., bijection). Indeed, a number of odor qualities were described by very specific rules. For example, 44% of the molecules described as *camphor* can be described by 3 rules physicochemical rules, with a very low rate of false positives (0.06%; molecules being described by the physicochemical rule, but not described perceptively as *camphor*). Similar patterns were observed for other qualities: e.g., *mint* (3 descriptive rules; 32% of the molecules described as *mint*; 0.06% of false positives), *ethereal* (3; 35%; 0%), *gassy* (3; 36%; 0.36%), *citrus* (3; 42%; 0.24%), *waxy* (3; 43%; 0%), *pineapple* (3; 48%; 0%), *medicinal* (3; 49%; 0.30%), *honey* (4; 54%; 0.06%), *sour* (3; 56%; 0.36%). Focusing on these qualities, this confirms, as stated above, that a universal rule cannot be defined for a given odorant property, in line with the extreme subtlety of our perception of smells. For example, looking in more details on the produced rules for Camphor (see Figure 16), it appears that one rule is mostly using topological descriptors, while the second rather uses chemical descriptors. The third rule has a combination of these two to fulfill the model.

8.5 Perspectives in neurosciences and chemistry

The present findings provide two important contributions to the field of neurosciences and chemo-sensation. First, although the SOR issue seems to be illusory for some odor qualities, our approach suggests that there exist descriptive rules for some qualities, and they also highlight the relevance of some

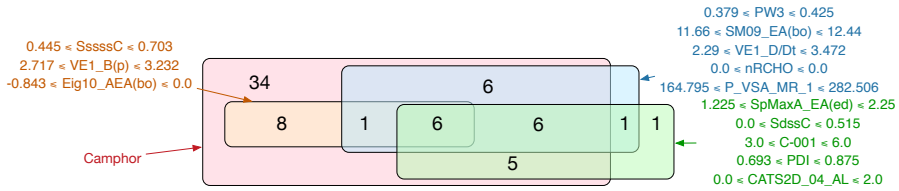


Fig. 16 Size of the support of three groups involving the *camphor* odor.

physicochemical descriptors (Sv, MW, etc.). Second, the present model confirms the lack of bijective (one-to-one) relationship between the odorant and the odor spaces and emphasizes that several sub-rules should be taken into account when producing structure-odor relationships. From these findings, experts in neurosciences and chemistry may generate the following new and innovative hypotheses in the field: (i) explaining inter-individual variability in terms of both behavioral and cognitive aspects of odor perception, (ii) explaining stability in odor-evoked neural responses and (iii) correlating the multiple molecular properties of odors to their perceptual qualities.

9 Conclusion

Motivated by a problem in neuroscience and olfaction, we proposed an original subgroup discovery approach to mine descriptive rules characterizing specifically subsets of class labels. For that matter, we revisited the *diverse subgroup set discovery* problem within the *exceptional model mining* framework: each subgroup can be evaluated in different *target subspaces*, and a quality measure helps to take into account the distribution of the labels overs the dataset. This measure is also effective in heuristic search, as the results are more diverse, both on the subgroup description and the target subspaces. We showed the effectiveness of this method through a deep set of experiments. Finally, the powerful interpretability of the results and the information they bring, can improve the knowledge about the complex phenomenon of olfaction.

Acknowledgements This research is partially supported by the *CNRS* (Préfute PEPS FASCIDO) and the *Institut rhônalpin des systèmes complexes* (IXXI).

References

1. S. Arctander. *Perfume and flavor materials of natural origin*, volume 2. Allured Publishing Corp., 1994.
2. M. Atzmüller and F. Lemmerich. Fast subgroup discovery for continuous target concepts. In *Foundations of Intelligent Systems, 18th International Symposium, ISMIS*, pages 35–44, 2009.
3. S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Min. Knowl. Discov.*, 5(3):213–246, 2001.

4. F. J. Berlanga, M. J. del Jesús, P. González, F. Herrera, and M. Mesonero. Multiobjective evolutionary induction of subgroup discovery fuzzy rules: A case study in marketing. In *Advances in Data Mining, Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, 6th Industrial Conference on Data Mining, ICDM 2006, Leipzig, Germany, July 14-15, 2006, Proceedings*, pages 337–349, 2006.
5. M. Boley, C. Lucchese, D. Paurat, and T. Gärtner. Direct local pattern sampling by efficient two-step random procedures. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 582–590, 2011.
6. G. Bosc, J.-F. Boulicaut, C. Raïssy, and M. Kaytoue. Anytime Discovery of a Diverse Set of Patterns with Monte Carlo Tree Search. *ArXiv e-prints*, Sept. 2016.
7. G. Bosc, M. Plantevit, J. Boulicaut, M. Bensafi, and M. Kaytoue. h(odor): Interactive discovery of hypotheses on the structure-odor relationship in neuroscience. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016*, pages 17–21, 2016.
8. C. Browne, E. J. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. P. Liebana, S. Samothrakis, and S. Colton. A survey of monte carlo tree search methods. *IEEE Trans. Comput. Intellig. and AI in Games*, 4(1):1–43, 2012.
9. L. Buck and R. Axel. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, 65(1):175–187, 1991.
10. C. J. Carmona, P. González, M. J. del Jesús, and F. Herrera. NMEEF-SD: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. *IEEE Trans. Fuzzy Systems*, 18(5):958–970, 2010.
11. C. A. de March, S. Ryu, G. Sicard, C. Moon, and J. Golebiowski. Structure-odour relationships reviewed in the postgenomic era. *Flavour and Fragrance Journal*, 30(5):342–361, 2015.
12. M. J. del Jesús, P. González, F. Herrera, and M. Mesonero. Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing. *IEEE Trans. Fuzzy Systems*, 15(4):578–592, 2007.
13. C. Delasalle, C. A. de March, U. J. Meierhenrich, H. Brevard, J. Golebiowski, and N. Baldovini. Structure-odor relationships of semisynthetic β-santalol analogs. *Chemistry & Biodiversity*, 11(11):1843–1860, 2014.
14. G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52, 1999.
15. L. Downar and W. Duivesteijn. Exceptionally monotone models - the rank correlation model class for exceptional model mining. In *2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015*, pages 111–120, 2015.
16. W. Duivesteijn, A. Feelders, and A. J. Knobbe. Different slopes for different folks: mining for exceptional regression models with cook's distance. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 868–876, 2012.
17. W. Duivesteijn, A. Feelders, and A. J. Knobbe. Exceptional model mining - supervised descriptive local pattern mining with complex target concepts. *Data Min. Knowl. Discov.*, 30(1):47–98, 2016.
18. W. Duivesteijn, A. J. Knobbe, A. Feelders, and M. van Leeuwen. Subgroup discovery meets bayesian networks – an exceptional model mining approach. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pages 158–167, 2010.
19. J. Fürnkranz, D. Gamberger, and N. Lavrač. *Foundations of Rule Learning*. Springer-Verlag, 2012.
20. B. Ganter and R. Wille. *Formal Concept Analysis*. Springer, 1999.
21. H. Grosskreutz and S. Rüping. On subgroup discovery in numerical domains. *Data Min. Knowl. Discov.*, 19(2):210–226, 2009.
22. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 1–12, 2000.

23. B. Kavsek, N. Lavrac, and V. Jovanoski. APRIORI-SD: adapting association rule learning to subgroup discovery. In *Advances in Intelligent Data Analysis V, 5th International Symposium on Intelligent Data Analysis*, pages 230–241, 2003.
24. M. Kaytoue, S. O. Kuznetsov, and A. Napoli. Revisiting numerical pattern mining with formal concept analysis. In *IJCAI*, pages 1342–1347, 2011.
25. M. Kaytoue, M. Plantevit, A. Zimmermann, A. Bendimerad, and C. Robardet. Exceptional contextual subgraph mining. *Machine Learning*, pages 1–41, 2017.
26. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, D. O. P. Consortium, G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, and P. Meyer. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017.
27. R. M. Khan, C.-H. Luk, A. Flinker, A. Aggarwal, H. Lapid, R. Haddad, and N. Sobel. Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *The Journal of Neuroscience*, 27(37):10015–10023, 2007.
28. W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. American Association for Artificial Intelligence, 1996.
29. L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006, Proceedings*, pages 282–293, 2006.
30. R. M. Konijn, W. Duivesteijn, M. Meeng, and A. J. Knobbe. Cost-based quality measures in subgroup discovery. *J. Intell. Inf. Syst.*, 45(3):337–355, 2015.
31. S. O. Kuznetsov. A fast algorithm for computing all intersections of objects in a finite semi-lattice. *Automatic Documentation and Mathematical Linguistics*, 27(5):400–412, 1993.
32. N. Lavrac, P. A. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In *Inductive Logic Programming, 9th International Workshop, ILP-99, Bled, Slovenia, June 24-27, 1999, Proceedings*, volume 1634 of *Lecture Notes in Computer Science*, pages 174–185. Springer, 1999.
33. D. Leman, A. Feelders, and A. J. Knobbe. Exceptional model mining. In *ECML/PKDD*, LNCS (5212), pages 1–16, 2008.
34. F. Lemmerich, M. Atzmueller, and F. Puppe. Fast exhaustive subgroup discovery with numerical target concepts. *Data Min. Knowl. Discov.*, 30(3):711–762, 2016.
35. F. Lemmerich, M. Becker, and M. Atzmueller. Generic pattern trees for exhaustive exceptional model mining. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, pages 277–292, 2012.
36. B. T. Lowerre. *The HARPY speech recognition system.* PhD thesis, Carnegie-Mellon Univ., Pittsburgh, PA. Dept. of Computer Science., 1976.
37. E. Loza Mencía and F. Janssen. Learning rules for multi-label classification: a stacking and a separate-and-conquer approach. *Machine Learning*, 105(1):77–126, 2016.
38. M. Meeng, W. Duivesteijn, and A. J. Knobbe. Rocsearch - an roc-guided search strategy for subgroup discovery. In *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, pages 704–712, 2014.
39. S. Moens and M. Boley. Instant exceptional model mining using weighted controlled pattern sampling. In *Advances in Intelligent Data Analysis XIII - 13th International Symposium, IDA 2014, Leuven, Belgium, October 30 - November 1, 2014. Proceedings*, pages 203–214, 2014.
40. P. K. Novak, N. Lavrac, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.
41. P. K. Novak, N. Lavrač, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.*, 10:377–403, 2009.
42. D. Rodríguez, R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz. Searching for rules to detect defective modules: A subgroup discovery approach. *Inf. Sci.*, 191:14–30, 2012.
43. G. Tsoumakas, I. Katakis, and I. P. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook, 2nd ed.*, pages 667–685. Springer, 2010.

44. G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414, 2011.
45. M. van Leeuwen and A. J. Knobbe. Non-redundant subgroup discovery in large and complex data. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011*, pages 459–474, 2011.
46. M. van Leeuwen and A. J. Knobbe. Diverse subgroup set discovery. *Data Min. Knowl. Discov.*, 25(2):208–242, 2012.
47. M. van Leeuwen and A. Ukkonen. Discovering skylines of subgroup sets. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013*, pages 272–287, 2013.
48. S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery, First European Symposium, PKDD '97, Trondheim, Norway, June 24-27, 1997, Proceedings*, pages 78–87, 1997.