

# Local Subgroup Discovery for Eliciting and Understanding New Structure-Odor Relationships

Guillaume Bosc<sup>1</sup>(✉), Jérôme Golebiowski<sup>3</sup>, Moustafa Bensafi<sup>4</sup>,  
Céline Robardet<sup>1</sup>, Marc Plantevit<sup>2</sup>, Jean-François Boulicaut<sup>1</sup>,  
and Mehdi Kaytoue<sup>1</sup>

<sup>1</sup> Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, 69621 Lyon, France  
guillaume.bosc@insa-lyon.fr

<sup>2</sup> Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205,  
69622 Lyon, France

<sup>3</sup> Université de Nice, CNRS, Institute of Chemistry, Nice, France

<sup>4</sup> Université de Lyon, CNRS, CRNL, UMR5292, INSERM U1028, Lyon, France

**Abstract.** From a molecule to the brain perception, olfaction is a complex phenomenon that remains to be fully understood in neuroscience. A challenge is to establish comprehensive rules between the physico-chemical properties of the molecules (e.g., weight, atom counts) and specific and small subsets of olfactory qualities (e.g., fruity, woody). This problem is particularly difficult as the current knowledge states that molecular properties only account for 30 % of the identity of an odor: predictive models are found lacking in providing universal rules. However, descriptive approaches enable to elicit local hypotheses, validated by domain experts, to understand the olfactory percept. Based on a new quality measure tailored for multi-labeled data with skewed distributions, our approach extracts the top- $k$  unredundant subgroups interpreted as descriptive rules  $description \rightarrow \{subset\ of\ labels\}$ . Our experiments on benchmark and olfaction datasets demonstrate the capabilities of our approach with direct applications for the perfume and flavor industries.

## 1 Introduction

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the century, the idea that modern, civilized human beings might do without being affected by odorant chemicals became outdated: the hidden, inarticulate sense associated with their perception, hitherto considered superfluous to cognition, became a focus of study in its own right and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., [2] 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 [4]. 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 [5]. 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. We create an interdisciplinary synergy between neuroscientists, chemists and data miners to the emergence of new hypotheses. Indeed, data-mining methods can be used to answer the SOR discovery problem, either through the building of predictive models or through rules discovery in pattern mining. One obstacle to this is that olfactory datasets are very complex (i.e., several thousand of dimensions, heterogeneous descriptors, multi-label, unbalanced classes, and non robust labelling) and, above all a lack of data-centric methods in neuroscience suitable for this level of complexity. The main aim of our study is to examine this issue by linking the multiple molecular characteristics of odorant molecule to olfactory qualities (fruity, floral, woody, etc.) using a descriptive approach (pattern mining). Indeed, a data science challenge was recently proposed by *IBM Research* and *Sage* [12]. Results suggest difficulties in the prediction of the data for *olfactory datasets* in general. The reason is that there is a strong inter- and intra-individual variability when individuals are asked about the quality of an odor. There are several explanations: geographical and cultural origins, each individual repertory of qualities (linguistic), genetic differences (determining olfactory receptors), troubles such as *anosmia* (see [3, 10]). It appears that designing pure predictive models remains today a challenge, because it depends on the individual’s genome, culture, etc. Most importantly, 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. For that, we develop a descriptive approach to make the results intelligible and actionable for the experts.

The discovery of (molecular) descriptions which distinguish a group of objects given a target (class label, i.e. odor quality(ies)) has been widely studied in AI, data mining, machine learning, etc. Particularly, supervised descriptive rules were formalized through subgroup discovery, emerging-pattern/contrast-sets mining, etc. [14]. In all cases, we face a set of objects associated to descriptions (which forms a partially ordered set), and these objects are related to one

or several class labels. The strength of the rule (SOR in our application) is evaluated through a quality measure (F1-measure, accuracy, etc.). The issues of multi-labeled datasets have been deeply studied in the state of the art [15]. However, to the best of our knowledge, most of existing methods to explore multi-label data are learning tasks. The existing descriptive approach known as Exceptional Model Mining (EMM) deals with multi-label data but it only considers them together, and not separately. Indeed, this method extracts subsets of objects (e.g., odorants) which distribution on all labels (e.g., odors) is statistically different (i.e., exceptional) w.r.t. the distribution of the entire set of objects. However, we aim to focus on subsets of few labels at a time. Moreover, the experts expect rules highlighting which values of features result in a subset of labels, and not only to extract relevant features for some labels as feature selection does. Our contributions are as follows:

- We explain the main problems of existing descriptive rule discovery approaches for dataset such as olfactory datasets, that are (i) multi-labeled with (ii) an unbalanced label distribution (i.e., a high variance in the labels occurrences).
- For (i), we consider the enumeration of pairs consisting of a description and a subset of labels (a variant of redescription mining [9]).
- For (ii), we propose a new measure derived from the F-score but less skewed by imbalance distribution of labels and that can dynamically consider the label distributions. We show this fact both theoretically and experimentally.
- We devise an algorithm which explores the search space with a beam-search strategy. It comes with two major issues that we jointly tackle: Finding the best cut points of numerical attributes during the exploration, also overcoming a redundancy among the extracted patterns.
- We thoroughly demonstrate the actionability of the discovered subgroups for neuroscientists and chemists.

The rest of the paper is organized as follows. We formally define the SOR discovery problem in Sect. 2 and we show why state-of-the-art methods are not adapted for this problem. We present our novel approach in Sect. 3 while the algorithmic details are given in Sect. 4. We report an extensive empirical study and demonstrate the actionability of the discovered rules in Sect. 5.

## 2 Problem Formulation

In this section, we formally define our data model as well as its main characteristics before introducing the problem of mining discriminant descriptive rules in these new settings. Indeed, we recall after the two most general approaches that can deal with our problem although only partially, namely *subgroup discovery* [14] and *redescription mining* [9]. We demonstrate this fact and highlight their weaknesses through an application example.

**Definition 1 (Dataset  $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ ).** Let  $\mathcal{O}$  and  $\mathcal{A}$  be respectively a set of objects (molecules) and a set of attributes (physicochemical properties). The value

**Table 1.** Toy olfactory dataset.

ID	MW	nAT	nC	Quality	ID	MW	nAT	nC	Quality
1	150.19	21	11	{Fruity}	4	152.16	23	11	{Fruity}
2	128.24	29	9	{Honey, Vanillin}	5	151.28	27	12	{Honey, Fruity}
3	136.16	24	10	{Honey, Fruity}	6	142.22	27	10	{Fruity}

domain of an attribute  $a \in \mathcal{A}$  is denoted by  $Dom(a)$  where  $a$  is said numerical if  $Dom(a)$  is embedded with an order relation, or nominal otherwise. Each object is described by a set of labels from the nominal set  $Dom(C)$  by the function  $class : \mathcal{O} \mapsto 2^{Dom(C)}$  that maps the olfactory qualities to each object.

**Running example.** Let us consider the toy olfactory dataset of Table 1 made of  $\mathcal{O} = \{1, 2, 3, 4, 5, 6\}$  the set of molecules IDS and  $\mathcal{A}$  the set of 3 of physicochemical attributes giving the molecular weight ( $MW$ ), the number of atoms ( $nAT$ ), and the number of carbon atoms ( $nC$ ). Each molecule is associated to one or several olfactory qualities from  $Dom(C) = \{Fruity, Honey, Vanillin\}$ . The assignments of an odor to a molecule is made by domain experts.

Real-life olfactory datasets, instances of this model, show specific characteristics: (i) **high dimensions**, (ii) **multi-label**, (iii) **unbalanced classes**, and (iv) **non-robust labeling**. Indeed, (i) the number of attributes is large, up to several thousands of physicochemical attributes and a hundred of labels ; (ii) a molecule takes several labels and (iii) the label distribution is highly unbalanced, i.e., with a high variance in the frequency with which the labels occur in the dataset. Odors like *fruity* (resp. *powdery*) are strongly over-represented (resp. under-represented) (see Fig. 1). Then, (iv) labels (odors) attached to each molecule are given by experts based on their own vocabulary. However there is both a high inter- and intra-individual variability concerning the perception of odors [12], the latter involving more than 400 genes encoding molecular receptors (whose expressions differ between people). Perception is subject to the context of the data acquisition phases (questionnaires), cultural elements, etc.

**Building an original 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 [12] 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](http://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 Fig. 1 (filled bars).

*Problem 1 (SOR Problem).* Given an olfactory dataset, the aim is to *characterize* and *describe* the relationships between the *physicochemical properties* of odorant molecules and their *olfactory qualities*.

**A data-science approach.** Answering this problem requires experts of different domains. The odor space is related to the study of olfaction in neuroscience, the understanding of the physicochemical space requires chemical skills, and finally, exploring jointly these two spaces requires data analysis techniques from computer science. In the latter, we cannot afford to use black box predictive models as we need intelligible patterns. Second, as olfactory datasets suffer from a poor label predictability, we cannot use global models to model the dataset but local models, i.e. subsets of data that are specific to some labels. These approaches are known as descriptive rule discovery methods [14], divided into two main trends: *subgroup discovery* and *redescription mining*. We introduce these methods and show their strengths and weaknesses to deal with our problem.

## 2.1 Subgroup Discovery

Subgroup Discovery (SD) has attracted a lot of attention for two decades under several vocables and research communities (subgroups, contrast sets, emerging patterns, etc.) [14,16]. The aim is to find groups of objects, called subgroups, for which the distribution over the labels is statistically different from that of the entire set of objects. A subgroup is defined (i) by its extent, i.e. the subset of objects it covers and (ii) by its intent, a description connecting restrictions on the attribute domains, such that the intent covers the extent. The intent can be defined on several languages, e.g. conjunctions of attribute domain restrictions.

**Definition 2 (Subgroup).** *The description of a subgroup is given by  $d = \langle f_1, \dots, f_{|\mathcal{A}|} \rangle$  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 covered by the description  $d$  is called the support of the subgroup  $\text{supp}(d) \subseteq \mathcal{O}$ . The set of all subgroups forms a lattice with a specialization/generalization ordering.*

**Definition 3 (Quality measure).** *The SD approach hence relies on a quality measure which evaluates the singularity of the subgroup within the population regarding a target class function: the class attribute. The choice of the measure depends on the dataset but also on the purpose of the application [8]. There are two main kind of quality measures: the first one is used with monolabeled dataset, e.g., the F-1 measure, the WRAcc measure, the Giny index or the entropy (the original SD [17]); and the second one is used with multilabeled dataset, e.g., the Weighted Kullback-Leibler divergence(WKL) as used in EMM [6]).*

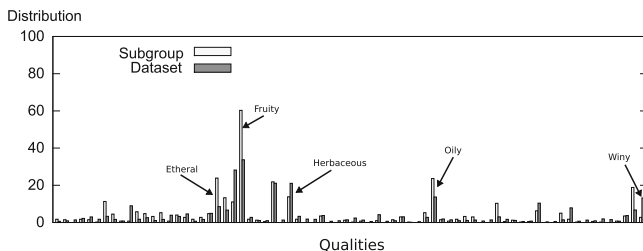


Fig. 1. Subgroup label distributions with WKL.

**Running example.** The support of the description  $d_1 = \langle MW \leq 151.28, 23 \leq nAT \rangle$  is  $\{2, 3, 5, 6\}$ . For readability, we omit a restriction  $f_i$  in a description if there is no effective restriction on the attribute  $a_i$ . The description  $d_2 = \langle MW \leq 151.28, 23 \leq nAT, 10 \leq nC \rangle$  is a specialization of  $d_1$  ( $d_1$  is a generalization of  $d_2$ ). Moreover, considering Table 1,  $WKL(d_1) = 4/6 \times ((3/4 \log_2 9/10) + (1/4 \log_2 3/2) + (3/4 \log_2 3/2)) = 0.31$ . The WRAcc measure of the descriptive rule  $d_1 \rightarrow \text{Honey}$  is  $WRAcc(d_1, \text{Honey}) = 4/6 \times (3/4 - 1/2) = 0.25$ .

**The SD problem.** Given a dataset  $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ ,  $minSupp$ ,  $\varphi$  and  $k$ , the objective is to extract the  $k$  best subgroups w.r.t. the measure  $\varphi$ , with a support cardinality higher than a given  $minSupp$ .

**Limits of SD addressing Problem 1.** The existing methods of SD either target only a single label at a time, or all labels together depending on the choice of the quality measure. In our application, it is required that a subgroup could characterize several labels at the same time. Only the WKL measure can achieve this goal [6]. However, it suffers of the curse of dimensionality: in presence of a large number of labels (74 odors in our experiments), the subgroups cannot characterize a small set of odors. This is shown with real data on Fig. 1: the distribution of labels for the full dataset and for the best subgroup are displayed: clearly, the subgroup is not characteristic of a few odors. We need thus to consider not only all the possible subgroups, but all label subsets for each subgroup. In other settings, this search space is actually considered by a method called *Redescription Mining* [9].

## 2.2 Redescription Mining

Redescription mining (RM) [9] aims at finding two ways of describing a same set of objects. For that, two datasets are given with different attributes but the same set of object ID. The goal is to find pairs of descriptions  $(c, d)$ , one in each dataset, where  $supp(c)$  and  $supp(d)$  are similar. The similarity is given by a Jaccard index between  $supp(c)$  and  $supp(d)$ . The closer to 1 the better the redescription. If we consider the first dataset as the chemicophysical attributes and the second as the labels, we can apply RM to find molecular descriptions and label sets that cover almost the same set of objects.

**Running example.** Let us consider the dataset of the Table 1 and the redescription  $r = (d_P, d_Q)$  with  $d_P = \langle MW \geq 150.19 \vee nAT = 27 \rangle$ ,  $d_Q = \langle Fruity \wedge (\neg Honey) \rangle$ . Thus,  $supp(d_P) = \{1, 4, 5, 6\}$  and  $supp(d_Q) = \{1, 4, 6\}$  and  $J(r) = \frac{3}{4} = 0.75$ . Note that RM allows an expressive language with negations and disjunctions for defining a description.

**The RM problem.** Given a dataset  $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ ,  $minSupp$  and  $k$ , the objective is to extract the  $k$  best redescriptions w.r.t. the Jaccard Index, with a support cardinality higher than  $minSupp$ .

**Limits of RM addressing Problem 1.** RM gives us an algorithmic basis to explore the search space composed of all pairs of subsets of objects and subsets of labels. However, the quality measure used in RM, the Jaccard index, does not fit exactly what Problem 1 expects. The Jaccard index is symmetric implying the discovery of almost bijective relationships. Yet, it is widely acknowledged that the structure-odor relationships are not bijective. Therefore, this measure is not relevant for unbalanced datasets, which is the case of olfactory datasets. Thus it is difficult to find descriptions related to an over-represented odor.

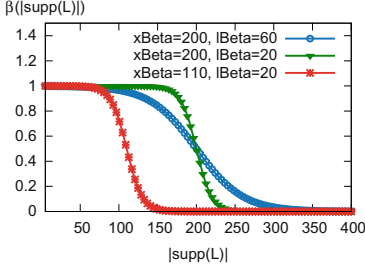
As a conclusion, answering Problem 1 can be achieved by exploring the search space of redescriptions of the form  $(d, L)$ , with  $d$  a description and  $L$  a subset of labels, using any quality measure from subgroup discovery (F1-measure, WRAcc, KWL, etc.). This however, does not take into account the *unbalanced classes* problem. We make this point explicit in the next section and propose a solution.

### 3 An Adaptive Quality Measure

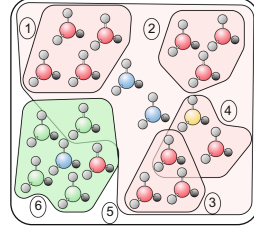
Existing discriminant descriptive rule methods cannot address Problem 1: the SD generic framework does not explore the correct search space whereas in RM the quality measure is not adapted for this problem. Problem 1 requires a data mining method that simultaneously explores both the description space and the search space of the odor labels. For that, we define *local subgroups*.

**Definition 4 (Local subgroup).** Given a dataset  $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ , a *local subgroup*  $(d, L)$  takes a description  $d$  characterizing a subset of few labels  $L \subseteq Dom(C)$  of the class attribute  $C$  with  $1 \leq |L| \ll |Dom(C)|$ . The support of a local subgroup is the support of its description:  $supp(d, L) = supp(d)$ . Note that  $supp(L) = \{o \in \mathcal{O} \mid L \subseteq class(o)\}$ .

The aim is to find out *local subgroups*  $(d, L)$  where the description  $d$  is characteristic of the subset of few olfactory qualities  $L \subseteq Dom(C)$ . For that, we develop a SD method, that simultaneously explores this double search space. This method relies on a adaptive quality measure that enables to evaluate the singularity of the local subgroup  $(d, L)$  only for the subset of labels  $L$  it targets. This measure is adaptive for each local subgroup, i.e., it is automatically adjusted according to the balance of the subset of labels in the dataset.



**Fig. 2.** The curves of  $\beta(|supp(L)|)$ .



**Fig. 3.** Necessity of an adaptive measure. (Color figure online)

**The original F-Score.** Complete surveys help understanding how to choose the right measure [8]. The generalized version of the WKL<sup>1</sup> considers the labels in the subset  $L \subseteq Dom(C)$  as independent and does not look for their co-occurrences. The WRAcc measure is a gain measure on the precision of the subgroup and totally ignores the recall. However, we are interested in a measure that considers both precision ( $P(d, L) = \frac{|supp(d) \cap supp(L)|}{|supp(d)|}$ ) and recall ( $R(d, L) = \frac{|supp(d) \cap supp(L)|}{|supp(L)|}$ ) of a local subgroup. The F-Score does it:

$$F(d, L) = (1 + \beta^2) \times \frac{P(d, L) \times R(d, L)}{(\beta^2 \times P(d, L)) + R(d, L)} \quad (1)$$

Indeed, objects are described by both attributes and class labels, so F-score quantifies both the precision and the recall of the support of the description w.r.t. the support of the class labels.

**The adaptive  $F_\beta$ .** However, olfactory datasets involve unbalanced labels, i.e. the distribution of the labels is quite different from each other: some are over-represented, and other are under-represented. Thus, we decided to adapt the F-Score to unbalanced datasets considering the original constant  $\beta$  as a variable of  $|supp(L)|$ : the higher  $|supp(L)|$ , the closer to zero  $\beta$  is (the precision in the F-Score is fostered), and the lower  $|supp(L)|$ , the closer to one  $\beta$  is (the F-Score becomes the harmonic mean of precision and recall). Formally, given two positive real numbers  $x_\beta$  and  $l_\beta$ , we define the  $F_\beta$  measure derived from Eq. 1 with  $\beta$  a variable of  $|supp(L)|$  as follows (see also Fig. 2)

$$\beta(|supp(L)|) = 0.5 \times \left( 1 + \tanh \left( \frac{x_\beta - |supp(L)|}{l_\beta} \right) \right) \quad (2)$$

Intuitively, for over-represented labels, 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 local subgroup  $(d, L)$  with plenty of non- $L$  odorants. In Fig. 3 the red odorants are over-represented in the dataset, but it is more interesting having the different local subgroups 1, 2, 3 and 4 with high precision, rather than a

<sup>1</sup> The generalized version of the WKL corresponds to the WKL measure restricted to the subset of labels  $L \subseteq Dom(C)$  of the local subgroup  $(d, L)$ .



single huge local subgroup 5 which precision is much lower. For odorants that are not over-represented, the measure considers both precision and recall: e.g., the local subgroup 5 is possible for the green molecules. The two real numbers  $x_\beta$  and  $l_\beta$  are set thanks to the characteristics of the dataset. In fact, due to the distribution  $\delta_C$  of the classes in the dataset, fixing  $x_\beta = E(\delta_L)$  and  $l_\beta = \sqrt{\sigma(\delta_L)}$ , where  $E(X)$  and  $\sigma(X)$  are respectively the average and the standard deviation of a random variable  $X$ , is sensible considering Problem 1.

**The local subgroup discovery problem.** Given a dataset  $\mathcal{D}(\mathcal{O}, \mathcal{A}, C, class)$ , the adaptive quality measure  $F_\beta$ , a minimum support threshold  $minSupp$  and an integer  $k \in \mathbb{N}^+$ , the aim is to extract the  $k$  best local subgroups  $(d, L)$  w.r.t. the quality measure  $F_\beta$  with  $1 \leq |L| \ll |Dom(C)|$ , such that  $supp(d, L) \geq minSupp$ .

**Running example.** Considering the dataset of Table 1, with  $x_\beta = 3$  and  $l_\beta = 1.4$ , let us discuss the local subgroup  $(d_1, \{Fruity\})$  with  $d_1 = \langle MW \leq 151.28, 23 \leq nAT \rangle$ . First,  $\beta(|supp(\{Fruity\})|) = 0.05$  since the *Fruity* odor is over-represented ( $|supp(\{Fruity\})| = 5$ ). We have that  $F_\beta(d_1, \{Fruity\}) = 0.75$  and it fosters the precision rather than the recall. Now if we consider the local subgroup  $(d_1, \{Honey, Fruity\})$ : since  $|supp(\{Honey, Fruity\})| = 2$ ,  $\beta(|supp(\{Honey, Fruity\})|) = 0.81$  from which it follows that  $F_\beta(d_1, \{Honey, Fruity\}) = 0.41$  because it considers both recall and precision.

## 4 Mining Local Subgroups

**Exploring the search space.** The search space of local subgroups is structured as the product of the lattice of subgroups and the lattice of label subsets, hence a lattice. Let  $X$  be the set of all possible subgroups, and  $C$  the set of labels, the search space is given by  $X \times 2^C$ . Each element of this lattice, called *node* or local subgroup hereafter, corresponds to a local subgroup  $(d, L)$ . Nodes are ordered with a specialization/generalization relation: the most general local subgroup corresponds to the top of the lattice and covers all objects, its set of labels is empty (or composed with labels that occur for –all– the objects). Each description  $d$  can be represented as a set of attribute restrictions: specializing a description is equivalent to add attribute domain restrictions (i.e. adding an element for nominal attributes, shrinking an interval to its nearest left or right value for a numerical attribute, see e.g. [11]).

Due to the exceptional size of the search space, we opt for a beam-search, starting from the most general local subgroup to the more specialized ones. This heuristic approach is also used in EMM and RM. It tries to specialize each local subgroup either by restricting an attribute or by extending subset of class labels with a new label it has also to characterize as long as the  $F_\beta$  score is improved. There are at most  $|Dom(C)| + \sum_{a_i \in \mathcal{A}} |a_i|(|a_i| + 1)/2$  possibilities to specialize each local subgroup: we can proceed up to  $|Dom(C)|$  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 width of the beam: the *beamWidth* best subgroups w.r.t. the quality measure). The search space is also pruned thanks to the anti-monotonic constraint on support.

**Finding the attribute split points.** When extending the description of a subgroup  $s = (d, L)$  for a numerical attribute, the beam search exploration looks for the best cut points that optimize the  $F_\beta$  score of the resulting subgroup. Since the value domain of a numerical attribute  $a$  is finite (at most  $|\mathcal{O}|$  different values), a naive approach would test all the possibilities to find the lower and the upper bounds for the interval that optimizes  $F_\beta$  ( $O(|\mathcal{O}|^2)$  complexity). Our approach, inspired by a state-of-the-art approach [7], only searches for promising cut points. We define  $r_i = \frac{|\{o \in \text{supp}(s) | a(o) = v_i, o \in \text{supp}(L)\}|}{|\{o \in \text{supp}(s) | a(o) = v_i, o \notin \text{supp}(L)\}|}$  for  $v_i \in \text{Dom}(a)$ . We say that a value  $v_i$  is a strict lower bound if  $r_i > 1$  and  $r_{i-1} \leq 1$ , and a value  $v_i$  is a strict upper bound if  $r_i > 1$  and  $r_{i+1} \leq 1$ . The algorithm searches for the best cut points among the strict lower and upper bounds.

**Table 2.** Characteristics of the datasets where  $|\mathcal{O}|$  is the number of objects,  $|\mathcal{A}|$  the number of attributes,  $|\mathcal{C}|$  the number of labels,  $M_1$  the average number of labels associated to an object, and  $M_2$ , *min*, *max* respectively the average, minimum and maximum number of objects associated to a label.

Dataset	$ \mathcal{O} $	$ \mathcal{A} $	$ \mathcal{C} $	$M_1$	$M_2$	<i>min</i>	<i>max</i>
$\mathcal{B}_1$	7395	243	159	2.4	111.7	51	1042
$\mathcal{D}_1$	1689	43	74	2.88	67.26	2	570
$\mathcal{D}_2$	1689	243	74	2.88	67.26	2	570

However this approach can return an empty set of cut points, especially for under-represented subsets of class labels. Experimentally, the beam search exploration stops very quickly and only over-represented label sets can be output. For that, we consider the *maxBeginningPoints* best lower bounds, i.e. value  $v_i$  such that  $0 < r_i \leq 1$ , as possible cut points when the original method of Fayyad et al. [7] does not return any result. By default we set *maxBeginningPoints* = 5.

**Mining diverse significant subgroups.** Generally, when a method mixes a beam search and a top-k approach, the issue of redundancy is clearly an important thing to deal with. The risk is to extract a set of top-k local subgroups where several subgroups are redundant, i.e. that share same restrictions or support. For that, we implement a process to avoid redundancy during the exploration. Before adding a local subgroup  $s$  in the top-k resulting set, we quantify the redundant aspect of  $s$  w.r.t. each current top-k local subgroup by comparing the restrictions involved in its description but also the support of these restrictions. Formally, we compute a penalty score  $pen(s_1, s_2) \in [0; 3]$  between two subgroups  $s_1$  and  $s_2$  by adding (i) the proportion of common attributes  $a_i$  involved in effective restrictions in both descriptions, and (ii) the values of the Jaccard index between the intervals  $[l_1, u_1]$  and  $[l_2, u_2]$  for each common attribute in the description,

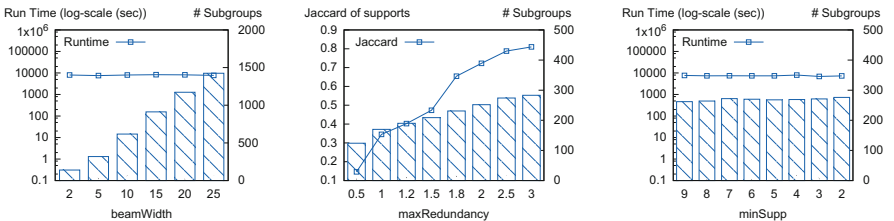
and (iii) the values of the Jaccard index between  $\text{supp}(s_1)$  and  $\text{supp}(s_2)$ . The algorithm only adds a new local subgroup if the penalty score with all other subgroups is less than the threshold  $\text{maxRedundancy}$ , and if the penalty score is greater than  $\text{maxRedundancy}$  the algorithm keeps the subgroup with the higher quality measure. By default, we fix  $\text{maxRedundancy}$  to 2.2.

Finally, extracted subgroups have to be statistically significant: considering a local subgroup  $(d, L)$ , the support of  $d$  and the support of  $L$  in the entire dataset have to be statistically meaningful. If we consider these distributions as independent, the probability that objects are included in both supports has to be low. To measure this, we compute the p-value: we test the distribution we face in the dataset against the null-model hypotheses.

## 5 Experiments

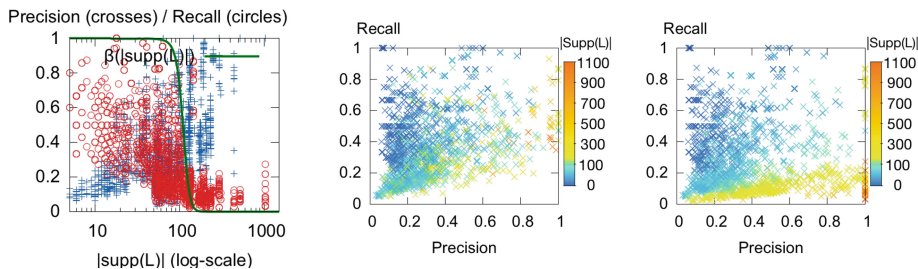
We experiment with the Bibtex dataset from the well-known MULAN<sup>2</sup> library for learning from multi-label datasets. The characteristics of this dataset are displayed in Table 2. The labels correspond to keywords the authors had chosen to their Bibtex entry. This Bibtex dataset is used to validate the method on both quantitative and qualitative sides because it does not require a deep expertise to interpret the results. We also used two real-world olfaction datasets. These datasets  $\mathcal{D}_1$  and  $\mathcal{D}_2$  have been derived from the dataset described in Sect. 2. Table 2 presents the characteristics of these datasets.

**Performance study.** To evaluate the efficiency of our algorithm, we consider the Bibtex dataset  $\mathcal{B}_1$  and the two olfaction datasets  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . Experiments were performed on a 3.10 GHz processor with 8 GB main memory running Ubuntu 14.04.1 LTS. We vary  $\text{minSupp}$ ,  $\text{beamWidth}$  and  $\text{maxOutput}$  separately and the non-varying parameters are fixed to  $\text{maxOutput} = 100$ ,  $\text{beamWidth} = 15$  and  $\text{minSupp} = 30$ . Surprisingly, in Fig. 4 (left), the runtime seems not to vary a lot when increasing the beam width. This is the same result when decreasing the minimum support threshold in Fig. 4 (right). This is due to the on-the-fly discretization method that is time-consuming. In

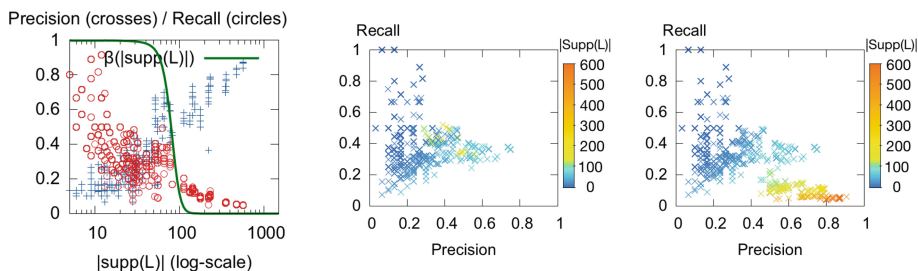


**Fig. 4.** The runtime and the number of output subgroups varying (left) the beam width, (middle)  $\text{maxRedundancy}$  and (right) the minimal support on  $\mathcal{D}_1$ .

<sup>2</sup> <http://mulan.sourceforge.net>.



**Fig. 5.** (left) The precision and the recall of subgroups  $(d, L)$  as a function of  $|supp(L)|$  on  $\mathcal{B}_1$  with the value of  $\beta(|supp(L)|)$ . The precision and the recall of the output subgroups  $(d, L)$  on  $\mathcal{B}_1$  according to  $|supp(L)|$  (color scale) using  $F_1$  (middle),  $F_\beta$  (right).



**Fig. 6.** The same experiments than those of Figs. 5 but on the dataset  $\mathcal{D}_1$ .

Fig. 4 (left and right), we observe that even if  $minSupp$  increases, the number of outputted subgroups is constant whereas when  $beamWidth$  increases, the number of extracted subgroups is higher. This is due to the avoiding redundancy task: when  $minSupp$  increases, the quality measure of the new generated local subgroups is improved, however, they may be redundant compared to other subgroups that are therefore removed. When  $beamWidth$  increases, the diversity is increased so the subgroups are less redundant. Figure 4 (middle) depicts the impact of our avoiding redundancy step. The lower  $maxRedundancy$ , the less similar the support of subgroups, the fewer extracted subgroups.

**Validating the adaptive F-measure.** Our choice to discover local subgroups  $(d, L)$  with an adaptive  $F_\beta$  score is well-suited for an olfactory dataset because a molecule is associated to a few olfactory qualities. For an experiment (the others highlight similar remarks), we have that 60.6 % of subgroups with  $|L| = 1$ , 33.8 % of subgroups with  $|L| = 2$  and 5.6 % of subgroups with  $|L| = 3$ . Figure 6 (left) depicts the impact of the factor  $\beta(|supp(L)|)$ . It displays for each extracted local subgroup  $(d, L)$  the precision and the recall of the descriptive rules  $d \rightarrow L$  as a function of  $|supp(L)|$ , with the curve of the factor  $\beta(|supp(L)|)$ . Clearly, it works as expected: the subgroups for which  $\beta(|supp(L)|)$  is close to 0 foster the precision rather than the recall, and the subgroup for which  $\beta(|supp(L)|)$  is close to 1 foster both recall and precision. Figure 6 (right) shows this point in

a different way: it displays the precision and the recall of each output subgroup ( $d, L$ ). A color code highlights the size of  $\text{supp}(L)$ : for over-represented labels, the precision is fostered at the expense of the recall whereas in other cases both precision and recall are fostered. Comparing to Fig. 6 (middle) which displays this result with the  $F_1$  score, we see that few output subgroups are relative to over-represented labels (the same applies for the Bibtex dataset  $\mathcal{B}_1$ , see Fig. 5).

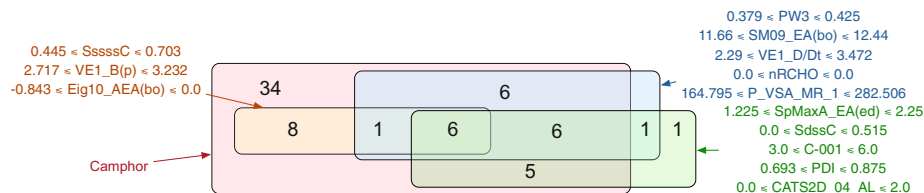
**Building a dataset for analyzing the olfactory percept.** We worked on our original dataset presented in Sect. 2. For this subsection, we derived 3 datasets by changing the following conditions. As our approach cannot handle 4,000 molecular descriptors: we filter out correlated attributes with the 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. We ran our algorithm on these three datasets with the combinations of different parameters: standard  $F_1$  score versus our adaptive measure  $F_\beta$ ;  $\text{minSupp} = 15$  (1 %) versus  $\text{minSupp} = 30$  (2 %): and finally, we experiment with three different thresholds for the  $\text{maxRedundancy}$  parameter (0.5, 1.5 and 2.5). All results are available at <http://liris.cnrs.fr/olfamining/>.

**Identification of relevant physicochemical attributes.** We consider the experiment on the dataset with 79 physicochemical properties, when we use the  $F_\beta$  score,  $\text{minSupp} = 30$ , and  $\text{maxRedundancy} = 2.5$ . A relevant information for neuroscientists and chemists concerns the physicochemical attributes that were identified in the descriptive rules. As showed in [13], 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,  $s_5 = \langle [27 \leq nOHs \leq 37] [6.095 \leq Sv \leq 7.871] [4 \leq N\% \leq 8] [25 \leq H\% \leq 28], \{Floral\} \rangle$  and  $s_6 = \langle [1 \leq nCsp2 \leq 1] [2.382 \leq TPSA(Tot) \leq 2.483] [4 \leq N\% \leq 10], \{Fruity\} \rangle$  are output subgroups. The quality measure of  $s_5$  is 0.91 with a precision of 0.91 and a low recall of 0.06. For  $s_6$ , its quality measure is up to 0.87, the same as its precision and its recall is 0.05. Each of these subgroups contains in its description the  $N\%$  attribute associated to a very low percentage, and  $s_5$  also includes the  $Sv$  attributes with a range of values that corresponds to its higher values. Note that, due to the  $F_\beta$  score, the recall of these subgroups is low because the odors *Fruity* and *Floral* are over-represented in the dataset. 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. For example,  $s_7 = \langle [5 \leq nCar \leq 6] [3.531 \leq Ui \leq 3.737] [224.43 \leq MW \leq 297.3], \{Musk\} \rangle$  with a quality measure of 0.46 (precision: 0.48, recall: 0.37) is about molecules with a molecular weigh between 224.43 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:  $s_8 = \{[3.453 \leq U_i \leq 3.691] [238 \leq MW \leq 297.3] [32 \leq nR = Cp \leq 87] [1 \leq nCsp2 \leq 6], \{Musk, Animal\}\}$ . This latter topological attribute is consistent with the presence of double bonds (or so-called  $sp^2$  carbon atoms) within most musky chemical structure, that provides them with a certain hydrophilicity.

**Providing relevant knowledge to solve a theoretical issue in the neuroscience of chemo-sensation.** We consider the experiment on the dataset with 615 physicochemical properties, when we use the  $F_\beta$  score,  $minSupp = 15$ , and  $maxRedundancy = 0.5$ . 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 Fig. 7), 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.

**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 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



**Fig. 7.** Size of the support of three groups involving the *camphor* odor.

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.

## 6 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, as well as an adaptive quality measure to be able to characterize both under- and over- represented label subsets. We implemented its algorithmic counterpart and experimented it with real olfactory datasets. The powerful interpretability of the results and the information they bring, can improve the knowledge about the complex phenomenon of olfaction. Applying such structure/odor model in a dedicated olfactory data-analytics platform will improve understanding of the effects of molecular structure on the perception of odorant objects (foods, desserts, perfumes, flavors), enabling product formulation to be optimized with respect to consumers' needs and expectations.

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

## References

1. Arctander, S.: *Perfume and Flavor Materials of Natural Origin*, vol. 2. Allured Publishing Corp., Carol Stream (1994)
2. Buck, L., Axel, R.: A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**(1), 175–187 (1991)
3. Castro, J.B., Ramanathan, A., Chennubhotla, C.S.: Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE* **8**(9), 09 (2013)
4. de March, C.A., Ryu, S., Sicard, G., Moon, C., Golebiowski, J.: Structure-odour relationships reviewed in the postgenomic era. *Flavour Fragr. J.* **30**(5), 342–361 (2015)
5. Delasalle, C., de March, C.A., Meierhenrich, U.J., Brevard, H., Golebiowski, J., Baldovini, N.: Structure-odor relationships of semisynthetic  $\beta$ -santalol analogs. *Chem. Biodivers.* **11**(11), 1843–1860 (2014)
6. Duivesteijn, W., Feelders, A., Knobbe, A.J.: Exceptional model mining - supervised descriptive local pattern mining with complex target concepts. *Data Min. Knowl. Discov.* **30**(1), 47–98 (2016)
7. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: *IJCAI* (1993)
8. Fürnkranz, J., Gamberger, D., Lavrač, N.: *Foundations of Rule Learning*. Springer, Heidelberg (2012)
9. Galbrun, E., Miettinen, P.: From black and white to full color: extending redescription mining outside the Boolean world. *Stat. Anal. Data Min.* **5**(4), 284–303 (2012)

10. Kaeppler, K., Mueller, F.: Odor classification: a review of factors influencing perception-based odor arrangements. *Chem. Senses* **38**(3), 189–209 (2013)
11. Kaytoue, M., Kuznetsov, S.O., Napoli, A.: Revisiting numerical pattern mining with formal concept analysis. In: *IJCAI*, pp. 1342–1347 (2011)
12. Keller, A., Vosshall, L., Meyer, P., Cecchi, G., Stolovitzky, G., Falcao, A., Norel, R., Norman, T., Hoff, B., Suver, C., Friend, S.: Dream olfaction prediction challenge (2015). [www.synapse.org/#!Synapse:syn2811262](http://www.synapse.org/#!Synapse:syn2811262). Sponsors: IFF, IBM Research, Sage Bionetworks and DREAM
13. Khan, R.M., Luk, C.-H., Flinker, A., Aggarwal, A., Lapid, H., Haddad, R., Sobel, N.: Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *J. Neurosci.* **27**(37), 10015–10023 (2007)
14. Novak, P.K., Lavrač, N., Webb, G.I.: Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.* **10**, 377–403 (2009)
15. Tsoumakas, G., Katakis, I., Vlahavas, I.P.: Mining multi-label data. In: *Data Mining and Knowledge Discovery Handbook*, 2nd edn., pp. 667–685 (2010)
16. van Leeuwen, M., Knobbe, A.J.: Diverse subgroup set discovery. *Data Min. Knowl. Discov.* **25**(2), 208–242 (2012)
17. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Komorowski, J., Zytkow, J. (eds.) *PKDD 1997*. LNCS, vol. 1263, pp. 78–87. Springer, Heidelberg (1997). doi:[10.1007/3-540-63223-9\\_108](https://doi.org/10.1007/3-540-63223-9_108)



Discovery Science

19th International Conference, DS 2016, Bari, Italy,

October 19-21, 2016, Proceedings

Calders, T.; Ceci, M.; Malerba, D. (Eds.)

2016, XXI, 492 p. 131 illus., Softcover

ISBN: 978-3-319-46306-3