h(odor): Interactive Discovery of Hypotheses on the Structure-Odor Relationship in Neuroscience

Guillaume Bosc[1(✉)], Marc Plantevit[1], Jean-François Boulicaut[1], Moustafa Bensafi[2], and Mehdi Kaytoue[1]

[1] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, 69621 Lyon, France
{guillaume.bosc,marc.plantevit,
jean-francois.boulicaut,mehdi.kaytoue}@liris.cnrs.fr
[2] Université de Lyon, CNRS, CRNL, UMR5292, INSERM U1028 Lyon, Lyon, France
moustafa.bensafi@liris.cnrs.fr

Abstract. From a molecule to the brain perception, olfaction is a complex phenomenon that remains to be fully understood in neuroscience. Latest studies reveal that the physico-chemical properties of volatile molecules can partly explain the odor perception. Neuroscientists are then looking for new hypotheses to guide their research: physico-chemical descriptors distinguishing a subset of perceived odors. To answer this problem, we present the platform *h(odor)* that implements descriptive rule discovery algorithms suited for this task. Most importantly, the olfaction experts can interact with the discovery algorithm to guide the search in a huge description space w.r.t their non-formalized background knowledge thanks to an ergonomic user interface.

1 Introduction

Olfaction, or the ability to perceive odors, was acknowledged as an object of science (Nobel prize 2004 [2]). The olfactory percept encoded in odorant chemicals contributes to our emotional balance and well-being. It is indeed agreed that the physico-chemical characteristics of odorants affect the olfactory percept [6], but 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? As only a part of the odorant message is encoded in the chemical structure, chemists and neuro-scientists are interested in eliciting hypotheses for the SOR problem under the form of human-readable descriptive rules: for example, $\langle MolecularWeight \leq 151.28, 23 \leq \#atoms \rangle \rightarrow \{Honey, Vanillin\}$. The discovery of such rules should bring new insights in the understanding of olfaction and has applications for Healthcare and the perfume and flavor industries.

Subgroup Discovery algorithms are able to discover such rules [7]. As olfaction datasets are composed of thousands of attributes, multi-labeled with a highly skewed distribution, an interactive mining of rules is interesting for experts that cannot formalize their domain knowledge, neither their mining preferences. Existing interactive subgroup discovery tools [3–5] can thus not be directly used due to the specificity of olfactory datasets. As such, we propose an original

B. Berendt et al. (Eds.): ECML PKDD 2016, Part III, LNAI 9853, pp. 17–21, 2016.
DOI: 10.1007/978-3-319-46131-1_4

platform, *h(odor)*, that enables to extract descriptive rules on physicochemical properties that distinguish odors through an interactive process between the algorithm and the neuroscientists.

2 System Overview

Input data and desired output. Our demo olfaction dataset is composed of 1,700 odorant molecules (objects) described by $1,500$ physicochemical descriptors [1] and are associated to several olfactory qualities (odors) among 74 odors given by scent experts. The data are represented in a tabular format (several CSV files). The physicochemical properties are numeric attributes and each olfactory quality is boolean. The goal is to extract subgroups $s = (d, L)$, i.e., descriptive rules, that covers a subset of molecules ($supp(s)$) where the description d over the physicochemical descriptors distinguishes a subset of odors L.

Algorithm sketch. The search space of subgroups is a lattice based on both the attribute space and the target space. The child $s' = (d', L')$ of a subgroup $s = (d, L)$ of the lattice is a specialization of s. This specialization consists of (i) restricting the interval of a descriptor in d, or (ii) adding a new odor to L. Since the search space growths exponentially with the number of descriptors and labels, a naive exploration (DFS or BFS) is not suitable. For that, we use the beam-search heuristic (BS). BS enables to proceed to a restricted BFS, i.e., for each level of the search space only a part of the subgroups are kept and put into the beam. Only the subgroups in the beam of the current level are explored in the next one [4]. The quality of a subgroup is evaluated by a measure. It adapts the F1-score by taking into account the label distribution for weighting the precision and recall.

System architecture. A core module (server) is contacted by a client (Web interface) to initiate the mining algorithm with the given parameters. This core module allows the user to interact/guide the algorithm exploration based on the likes/dislikes of the user (Fig. 1).

Core Module. This is the back-end of the *h(odor)* application. Based on *NodeJS*, the *Core Module* is the gateway between the user and the algorithm: it is in charge of the interaction. For that, JSON data are sent to and received from the *SD Algorithm* through sockets thanks to a dedicated communication process. Moreover, this module controls the UI to display results extracted from the *SD Algorithm* and collects the user preferences (like/dislike).

User Interface (UI). The front-end of the application, based on *Bootstrap* and *AngularJS*, enables the user to select the parameters of the *SD Algorithm* and to run it. Once the subgroups of the first level of the *beam search* are extracted (the algorithm

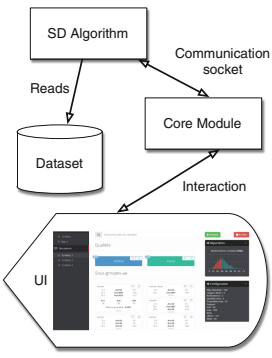

Fig. 1. System architecture

is paused waiting the user preferences), the UI displays these subgroups and the user can like/dislike some of them: the liked subgroups are forced to be within the beam for the next step, and the disliked subgroups are removed from the beam. When the algorithm finishes, the UI displays the results.

3 Use Case: Eliciting Hypotheses for the Musk Odor

We develop a use case of the application as an end user, typically a neuroscientist or a odor-chemist that seeks to extract descriptive rules to study the Structure-Odor Relationships. The application is available online with a video tutorial supporting this use case http://liris.cnrs.fr/dm2l/hodor/. In this scenario, we proceed in the following steps, knowing that the expert wishes to discover rules involving at least the *musk* odor.

1- Algorithms, parameters and dataset selection. In the *Algorithms* section of the left hand side menu, the user can choose the exploration method and its parameters. In this use case, we consider the ELMMut algorithm. This algorithm implements a beam search strategy to extract subgroups based on a quality measure. We plan to add new/existing algorithms and subgroup quality measures. Once the exploration method is chosen, we have to select the olfactory dataset as introduced in the previous section, and choose to focus on the *musk* odor. Considering our use case, we decided to set the size of the beam to 50 (the exploration is quite large enough) and the minimum support threshold to 15 (since $|supp(Musk)| = 52$, at least the subgroups have to cover 30 % of the musk odorants). Other parameters are fixed to their default value.

2- Interactive running steps. When the datasets and the parameters have been fixed, the user can launch the mining task clicking on the *Start mining* button. When the first step of the *beam search* is finished, the *SD Algorithm* is paused and the subgroups obtained at this step are displayed to the user. The interaction view in the front-end presents the olfactory qualities involved at this level of the exploration (see Fig. 2). Each subgroup is displayed in a white box with the current descriptive rule on the physicochemical descriptors and some quantitative measures. For each subgroup box, the user can select in the top right corner if he likes/dislikes this subgroup. For example, at the first step, the application displays the subgroups extracted at the first level for the *Musk* odor. As it is a known fact in chemistry that the *musk* odor involves large molecules, we *like* the subgroup which description is $d = [238.46 \leq MW \leq 270.41]$. After that, we keep on exploring by clicking the *Next* button. Another interactive step begins, but the expert has no particular opinion so he can jump to the *next* level. Once the algorithm finished (the quality measures cannot improve), we can study the table of results. For example, the description of one of the best extracted subgroups s is: $[238.46 \leq MW \leq 270.41][-0.931 \leq Hy \leq -0.621][2.714 \leq MLOGP \leq 4.817][384.96 \leq SAtot \leq 447.506][0 \leq nR07 \leq 0][0 \leq ARR \leq 0.316][1 \leq nCsp2 \leq 7]$ that involves large odorants. Moreover, according to the experts, this latter topological descriptor 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. The goal of the *h(odor)* application is to confirm knowledge and to elicit new

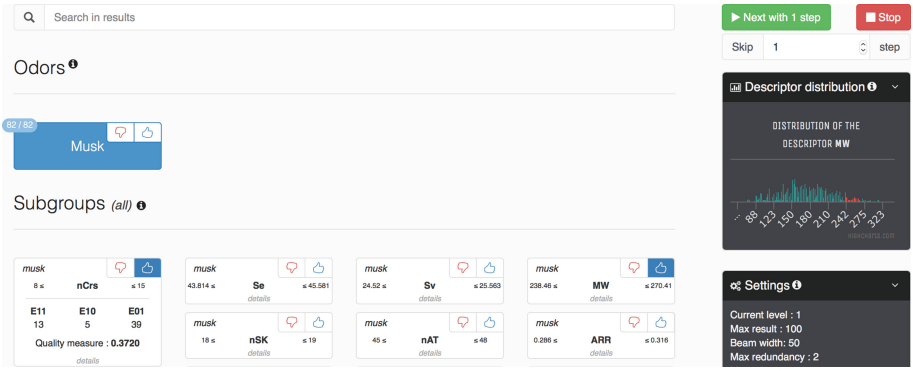

Fig. 2. The interaction view of the application. For each step of the beam search, the algorithm waits for the user's preferences (like/dislike). The subgroups are displayed into white boxes. On the right part, complementary information is displayed: part of value domain of a chosen restriction on a descriptor, and parameters of the run.

hypotheses for the SOR problem. In the case of s, the neuroscientists are interested in understanding why these descriptors (excepted the Molecular Weight) are involved in the $Musk$ odor.

Learning user preferences. Besides, the *h(odor)* application enables to save all the choices taken by the different users. Indeed, the application archived all the actions the users did into log files. The goal here is to use these log files to learn user preferences, not only for a single run of the algorithm [3] but for all experiments performed by the users. This kind data (choices made by experts) is hard to collect by simply asking experts and will be explored in future work.

Acknowledgments. The authors thank Florian Paturaux, Sylvio Menubarbe and Pierre Houdyer for helping developing the prototype. This research is partially supported by the *Institut rhônalpin des systémes complexes* (IXXI) and by the *Centre National de Recheche Scientifique* (Préfute PEPS FASCIDO, CNRS).

References

1. Arctander, S.: Perfume and Flavor Materials of Natural Origin, vol. 2. Allured Publishing Corp., USA (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. Dzyuba, V., van Leeuwen, M., Nijssen, S., Raedt, L.D.: Interactive learning of pattern rankings. Int. J. Artif. Intell. Tools **23**(6), 1–31 (2014)
4. Galbrun, E., Miettinen, P.: Siren: an interactive tool for mining and visualizing geospatial redescriptions. In: KDD, pp. 1544–1547 (2012)

5. Goethals, B., Moens, S., Vreeken, J.: MIME: a framework for interactive visual pattern mining. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011. LNCS (LNAI), pp. 634–637. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23808-6_45
6. de March, C.A., Ryu, S., Sicard, G., Moon, C., Golebiowski, J.: Structure-odour relationships reviewed in the postgenomic era. Flavour Fragrance J. **30**(5), 342–361 (2015). http://dx.doi.org/10.1002/ffj.3249
7. 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)