Co-occurrence-based clustering of odor descriptors for predicting structure-odor relationship

Conference Paper · May 2019

DOI: 10.1109/ISOEN.2019.8823446

3 authors:

Chuanjun Liu
Kyushu University
91 PUBLICATIONS **1,661** CITATIONS
SEE PROFILE

Liang Shang
Chinese Academy of Sciences
22 PUBLICATIONS **477** CITATIONS
SEE PROFILE

Kenshi Hayashi
Kyushu University
239 PUBLICATIONS **2,877** CITATIONS
SEE PROFILE

CO-OCCURRENCE-BASED CLUSTERING OF ODOR DESCRIPTORS FOR PREDICTING STRUCTURE-ODOR RELATIONSHIP

Chuanjun Liu[1,2], Liang Shang[2], Kenshi Hayashi[2]
[1] Research Laboratory, U.S.E. Co., Ltd., 22-10 Ebisu 4-Chome, Shibuya-ku, Tokyo 150-0013, Japan
[2] Graduate School of Information Science and Electrical Engineering, Kyushu University, 744, Motooka, Nishi-ku, Fukuoka, 819-0395, Japan

ABSTRACT

One problem of machine-learning-based prediction of structure-odor relationship is that odorant molecules are usually labeled with ambiguous descriptors when they are collected from different sources. This study focused on the clustering of the odor descriptors by text mining approaches as well as the prediction of newly established labels from physicochemical parameters of the classified odorant molecules. An odor database was established by web scraping and transferred to a document-term matrix including 4011 odorants and 100 odor descriptors. The clustering of the odor descriptors was carried out by using different co-occurrence matrix and clustering approaches. A hierarchical cluster analysis combined with a co-occurrence probability distribution matrix has shown good results in the descriptor clustering. The attribute labels of each class were established and then predicted from physicochemical parameters of the classified odorants by using random forest model. An average accuracy higher than 82.42% was obtained, indicating the effectiveness of the proposed approaches for predicting structure-odor relationship.

Index Terms— structure-odor relationship, text mining, machine learning, odor descriptors, physicochemical parameters, clustering, prediction

1. INTRODUCTION

Although understanding the relationship between the molecular structure of odorants and their olfactory perception has long remained a major challenge, the data-driven approaches rapidly developed in recent years has made it possible to predict the olfactory perception from their physicochemical properties [1,2,]. Such kind of predictions may not only deepen our understanding on the olfactory perception of odorant molecules, but also explore new applications that cannot be realized before. For example, in our previous work we have introduced a machine-learning-based gas chromatography-olfactometry (GCO), in which the olfactometric detection can be done by a machine classifier, and thus the human panelist might be no longer needed [3,4].

The previous study was carried out using a dataset with limited odorants and odor descriptors. For the practical application of structure-odor relationship prediction, the concept should be proofed by data on a large scale. The problem of sample size can be solved by collecting data from open or commercial sources as many researches have done [5,6]. However, there exists an phenomenon in the collected database that the odorant molecules are usually noted by descriptors with ambiguous, vague and unintelligible meaning. This may make it difficult to correctly label the odorant molecules for the machine learning, and thus decreases the prediction accuracy of models.

In this study, we extend the perception prediction of odorant molecules by using an enlarged database that is collected from different sources. Differently from the single descriptor-labeled prediction in the previous work, the odor descriptors are firstly clustered based on their co-occurrence relationship in a document term matrix by various classification approaches. The clustered odor descriptors with semantic similarity are endowed with new attribute labels, which is used to group odorant molecules again. The predict of the group labels is carried out by a random forest model from physicochemical parameters of the odorant molecules.

2. MATERIALS AND METHODS

Raw data including the CAS number and odor descriptors was obtained by web scraping from different websites [7]. R software was used in data process as well as model prediction. Ambiguous, lengthy and irrelevant expressions were sorted and integrated by both semi-automatic and manual methods. Compound descriptor words (or sentences) were split and only single adjective word was used to describe the odorant molecules (for example, “coffee-like” to “coffee”, “roasted meat” to “roasted” and “meaty”). The olfaction-unrelated vocabularies were also excluded (for example, “sharp”, “rich” and “odorless”). The raw data included ~ 5000 odorants and ~ 500 descriptors. In order to ensure enough sample size for co-occurrence analysis as well as the model prediction, we only counted the descriptors with top 100 frequency. Therefore, the final document-term matrix consisted of 4011 molecules and 100 descriptors.

A co-occurrence matrix (100×100 dimensions) of the 100 descriptors was established from the document-term matrix. Two major approaches were used in the descriptor clustering: pairwise-based clustering and similarity-based clustering.

The principle of pairwise-based clustering is that if terms ω_1 and ω_2 co-occur frequently, they are considered to be in the same cluster. The principle of similarity-based clustering is that if term ω_1 and ω_2 have similar distribution of co-occurrence with other terms, they are considered to be in the same cluster [8,9]. The pairwise-based clustering was carried out by using Jaccard index matrix (we called it J-matrix) and the similarity-based clustering was carried out by using co-occurrence probability distribution matrix (we called it D-matrix). Both hierarchical (hclust) and non-hierarchical (k-means) approaches were applied in the descriptor clustering based on the above two matrix. Odorant molecules labeled with odor descriptors belonging to the same cluster were grouped together with new attribute labels for the next prediction.

The physicochemical parameters of the odorant molecules were calculated by Dragon cheminformation software (Ver 7.0, Kode, Italy). Parameters with "NA" (not applicable), as well as those zero values that cannot be used in data scaling, were removed from the dataset, which finally afforded a matrix with 4011 odorants and 1780 parameters. These parameters include molecular weight, number of atoms, aromatic ratio, topological charge index, and so on (for detailed information please refer to [3]). Principal component analysis (PCA) was used in the feature extraction from the molecular parameters. To avoid loss of characteristic information from the original dataset, we selected PCs with accumulative contributions of 99.99%, which included approximately 200 PCs.

Since the data of the labeled classes is a typical imbalanced dataset, synthetic minority oversampling technique (SMOTE) [10] was used to balance both the minor and major samples according to actual needs. After the sample balancing, The SMOTE samples were then divided into training and test sets with a ratio of 4:1 by using the createDataPartition function of the caret package. The predict of classified attribute labels was carried out by using the package 'randomForest', in which parameters such as the number of trees and the number of variables are optimized automatically. A 5-fold cross-validation approach was applied in the sample splitting to confirm the model generalization. The prediction accuracy was evaluated by a mean value of 5 validations with the standard deviation.

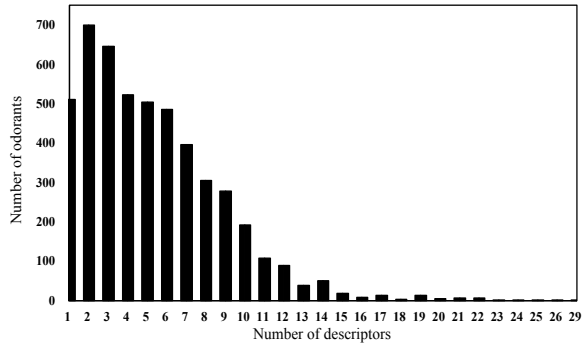

Figure 1 Number distribution of odorants by descriptors.

3. RESULTS AND DISCUSSION

Figure 1 shows the frequency distribution of odor descriptors by the number of odorant molecules in the raw database. Table 1 shows the number of odorants by the top 100 descriptors. It can be seen that the top one descriptor (fruity) is almost associated by one third of the odorant molecules (~1400). An odorant molecule (herein methyl eugenol) is noted by descriptors with the maximum number of 29. This phenomenon that one odorant is represented by so many descriptors and one descriptor is shared by so many odorants makes it difficult to precisely label the odorant molecules. This will influence the feature extraction from the molecular structure as well as the following machine learning. Therefore, it is necessary to cluster the odor descriptors to obtain more precise labeling information. Moreover, The minimum number of odorant molecules for the top 100 descriptors is larger than 58, which may ensure the sample size of the prediction models.

We firstly compared the clustering effect of the descriptors by PCA when different data matrix (J-matrix and D-matrix) and algorithms (hclust and k-means) were applied. Figure 2 shows the visualized results of four combinations: J-hclust, J-kmean, D-hlcust and D-kmean, respectively. It is noticed that the cumulative contribution rates of the two principle components (calculated by the sum of Dim 1 and Dim 2) in the case of D-matrix (24.5%) is much higher than that of J-matrix (8.0%). These results indicate that the co-occurrence probability distribution matrix (D-matrix) may has much better effect than the Jaccard index matrix to extract the similarity of odor descriptors. Additionally, in the case of D-matrix, the hierarchical classification (hclust) shows less overlapping between different descriptor clusters. It is then suggested that the D-hclust approach is much better for the descriptor clustering.

Table 1 Number of odorants by the top 100 descriptors

fruity	1402	roasted	195	cheese	111	lily	72
sweet	1244	dry	191	tobacco	111	natural	72
green	1235	honey	186	onion	110	smoky	72
floral	999	creamy	185	terpenic	105	soap	70
woody	896	wine	176	anise	103	dairy	69
herbal	802	camphor	175	clean	103	grassy	69
spicy	606	berry	171	warm	103	alliaceous	65
fresh	492	pineapple	166	pungent	101	muguet	65
fatty	486	caramel	165	coconutty	100	cortex	64
citrus	431	ethereal	158	musk	100	plum	63
balsamic	429	phenolic	154	burnt	98	bitter	62
waxy	417	orange	149	lemon	97	cinnamon	62
rose	376	aldehydic	145	peach	96	geranium	62
earthy	332	leaf	142	pepper	95	tea	62
nutty	328	grape	137	almond	94	milk	61
sulfurous	279	amber	130	mushroom	92	orris	61
oily	274	vanilla	130	cherry	87	rummy	61
tropical	269	metallic	122	chocolate	87	hyacinth	60
minty	244	melon	121	violet	87	lavender	60
pine	237	pear	117	apricot	86	clove	59
apple	216	coffee	115	cedar	83	coumarin	59
meaty	216	animal	114	medicinal	83	peel	59
vegetable	208	jasmine	112	garlic	79	raspberry	58
musty	195	banana	111	cocoa	76	strawberry	58
powdery	195	butter	111	hay	75	weedy	58

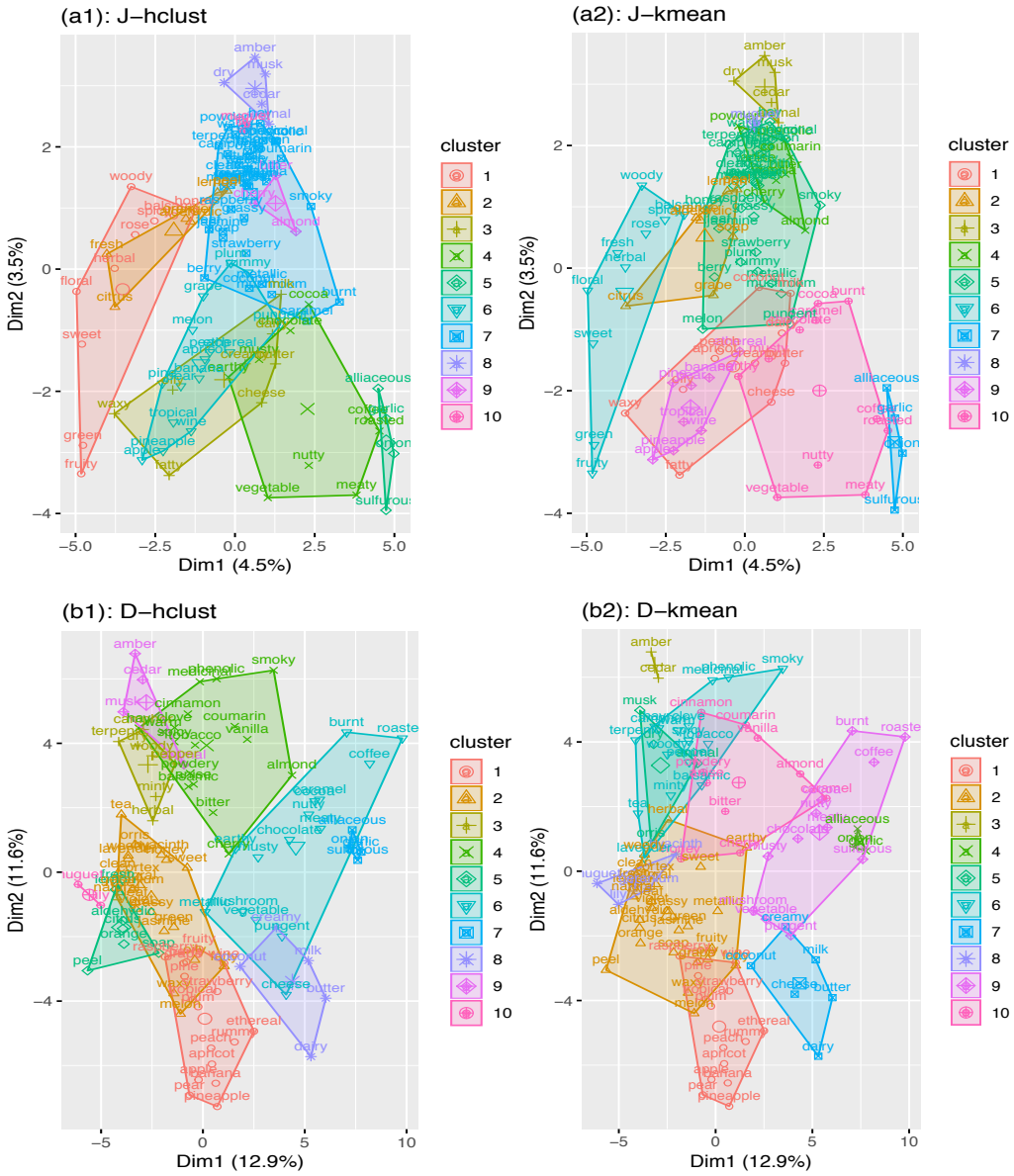

Figure 2 Visualized results of odor descriptors by different data matrix (J-matrix and D-matrix) and different clustering approaches (hierarchical and k-means).

Figure 3 shows the dendrogram of the 100 descriptors based on the D-hclust approach. The number of classes was tentatively fixed as ten in this plot. It can be seen that the descriptors with semantic similarity are almost divided into the same class. For example, alliaceous, sulfurous, onion and garlic are clustered into C2. The descriptors related to tropic fruit are clustered together in C7. Therefore, it is possible to annotate the odorant communities with new attribute labels (this is another topic and not discussed in this work). The rearrangement of odorant molecules may hold more semantic commonality, which will be helpful for the feature extraction from physicochemical parameters.

Table 2 lists the prediction results of the above 10 classes by using a random forest model for the test subset (20% of the SMOTE samples). The highest mean prediction accuracy (96.24%) is observed for the class of C10, while the minimum accuracy (69.66%) is observed for the class of C6. The average accuracy of the 10 classes is about 82.42%. Relatively low accuracy was observed for classes with diversified descriptors and more branches in the dendrogram. It is found that if the class number increases (that can be done by adjusting the cut-off distance), both the minimum and average accuracies will increase further. For example, when the class number is fixed as 16, the obtained minimum and average accuracies are 74.22% and 85.32%, respectively (data not shown). This result suggests the determination of appropriate class number might be an important issue for the prediction.

Table 2 Model parameters and prediction accuracies of 10 odorant classes.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
P	356	283	1293	309	1319	1691	1163	3199	951	89
N	3655	3728	2718	3702	2692	2320	2848	812	3060	3922
over	400	400	100	400	100	25	100	200	200	1000
under	120	120	200	120	200	500	200	150	150	120
P_SMOTE	1780	1415	2586	1545	2638	2113	2326	2436	2853	1068
N_SMOTE	1708	1358	2586	1483	2638	2110	2326	2436	2853	979
TP*	328	256	375	273	365	308	360	382	461	186
FP*	28	27	142	36	162	114	105	105	109	9
FN*	33	26	118	30	102	142	106	83	106	6
TN*	308	245	399	266	425	280	359	404	464	207
Acc%	89.31	90.70	75.42	88.77	74.92	69.66	76.40	81.20	81.55	96.24
SD%	0.89	0.66	2.46	0.62	0.63	1.47	0.96	1.07	0.38	0.50

P/N means positive and negative sample of individual class; over/under: parameters used to expand the minority sample and reduce the majority samples in SMOTE; P_SMOTE/N_SMOTE: positive and negative samples after balanced by SMOTE. TN: True negative, FP: False positive, FN: False negative, TP: True positive (*: the value is an example of one cross-validation); Acc is the mean prediction accuracy of 5-fold cross-validation. SD is the standard deviation.

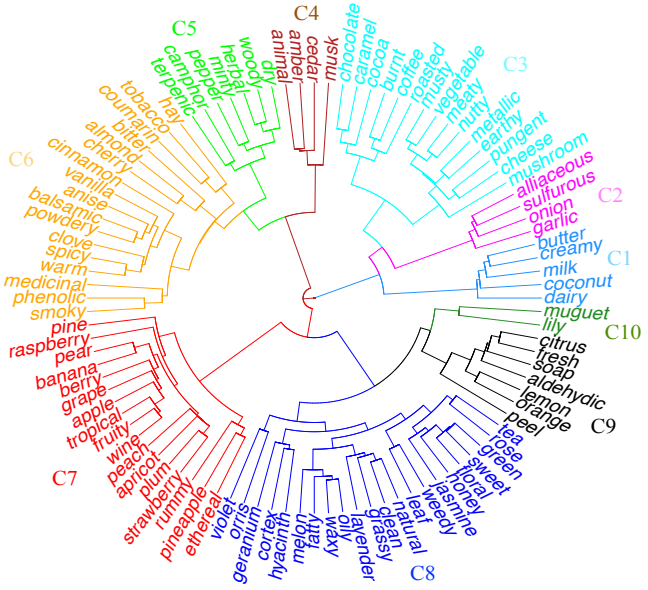

Figure 3 Dendrogram of odor descriptors by D-hclust.

4. CONCLUSION

Text mining approaches were introduced in the classification of odor descriptors and odorant molecules for predicting structure-odor relationship. Compared with our previous prediction based on single descriptor label, the descriptor classification enlarged the range of prediction because more descriptors and odorants were considered. We suggested that the classification of odor descriptors might help to the feature extraction from both the odor descriptors and the molecular structures, and thus to improve the prediction accuracy of machine-learning-based olfactory perception.

5. REFERENCES

[1] K. Kaeppler, F. Mueller, "Odor classification: a review of factors influencing perception-based odor arrangement", *Chemical Senses*, 38, 189-209 (2013)

[2] B. Auffarth, "Understanding smell – The olfactory stimulus problem", *Neuroscience and Biobehavioral Reviews*, 37, 1667-1669 (2013)

[3] L. Shang, C. Liu, Y. Tomiura, K. Hayashi, "Machine-learning-based olfactometer: prediction of odor perception from physicochemical features of odorant molecules" *Analytical Chemistry*, 89,11999-12005 (2017)

[4] L. Shang, C. Liu, Y. Tomiura, K. Hayashi, "Odorant clustering based on molecular parameter-feature extraction and imaging analysis of olfactory bulb odor maps", *Sensors and Actuators B: Chemcial*, 255, 508-518 (2018)

[5] R. Kumar, R. Kaur, B. Auffarth, A.P. Bhondekar, "Understanding the odor spaces: A step towards solving olfactory stimulus-percept problem", *PLOS ONE*, 10(10):e0141263(2015)

[6] A. Tromelin, C. Chabanet, K. Audouze, F. Koensgen, "Multivariate statisrical analysis of a large odorants database aimed at revealing similarities and links between odorants and odors" *Flavour Fragrance Jouranl*, 33,106-126 (2018)

[7] OdorMapOB: https://senselab.med.yale.edu/odormapdb/; SuperScent: http://bioinf-applied.charite.de/superscent/; GoodScents:http://www.thegoodscentscompany.com; Flavornet:http://www.flavornet.org/index.html; Sigam-Aldrich: https://www.sigmaaldrich.com/industries/flavors-and-fragrances.html

[8] Y. Matsuo, M. Ishizuka, "Keyword extraction from a document using word co-occurrence statistical information" *International Journal on Artificial Intelligence*, 13 (1), 157-169 (2002)

[9] N. Takahashi, A. Okabe, "Word clustering based on co-occurrence information in English GIS textbooks", *Theory and Application of GIS*, 15(2), 129-136 (2007)

[10] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, 16, 321-357 (2002)