УДК 543.63; 543.544; 004.8

Gradient boosting for the prediction of gas chromatographic retention indices

© 2019 Matyushin D.D., Sholokhova A.Yu., Buryak A.K.

A.N. Frumkin Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Moscow

Received 10.11.2019

DOI: 10.17308/sorpchrom.2019.19/2223

The estimation of gas chromatographic retention indices based on compounds structures is an important problem. Predicted retention indices can be used in a mass spectral library search for the identification of unknowns. Various machine learning methods are used for this task, but methods based on decision trees, in particular gradient boosting, are not used widely. The aim of this work is to examine the usability of this method for the retention index prediction. 177 molecular descriptors computed with Chemistry Development Kit are used as the input representation of a molecule. Random subsets of the whole NIST 17 database are used as training, test and validation sets. 8000 trees with 6 leaves each are used. A neural network with one hidden layer (90 hidden nodes) is used for the comparison. The same data sets and the set of descriptors are used for the neural network and gradient boosting. The model based on gradient boosting outperforms the neural network with one hidden layer for subsets of NIST 17 and for the set of essential oils. The performance of this model is comparable or better than performance of other modern retention prediction models. The average relative deviation is ~3.0%, the median relative deviation is ~1.7% for subsets of NIST 17. The median absolute deviation is ~34 retention index units. Only non-polar liquid stationary phases (such as polydimethylsiloxane, 5% phenyl 95% polydimethylsiloxane, squalane) are considered. Errors obtained with different machine learning algorithms and with the same representation of the molecule strongly correlate with each other.

Keywords: gas chromatography, retention index, machine learning, gradient boosting

Градиентный бустинг для предсказания газохроматографических индексов удерживания

© 2019 Матюшин Д.Д., Шолохова А.Ю., Буряк А.К.

Институт физической химии и электрохимии им. А.Н. Фрумкина РАН, Москва

Оценка газохроматографических индексов удерживания исходя из структур молекул является важной задачей. Предсказанные индексы удерживания могут быть использованы при идентификации неизвестных соединений посредством поиска по масс-спектральным базам данных. Разнообразные методы машинного обучения используются для этой задачи, однако, методы, основанные на деревьях решений, в частности градиентный бустинг (gradient boosting), не часто используются для этой цели. Цель этой работы – изучить возможность использования этого метода для предсказания индекса удерживания. 177 молекулярных дескрипторов, рассчитанных с помощью Chemistry Development Kit, используются в качестве входного представления молекулы. Случайные подмножества всей базы данных NIST 17 используются в качестве наборов данных для обучения, тестирования и валидации. Используется 8000 деревьев решений, имеющих по 6 листьев (конечных узлов) каждое. Нейронная сеть с одним скрытым слоем, состоящим из 90 скрытых нейронов, используется для сравнения. И нейронная сеть, и градиентный бустинг используются с одним и тем же набором молекулярных дескрипторов и одними и теми же наборами данных. Модель, основанная на градиентном бустинге, превосходит нейронную сеть с одним скрытым слоем для подмножеств NIST 17 и для набора эфирных масел. Основанная на градиентном бустинге модель сопоставима или даже превосходит по точности другие современные модели предсказания индексов удерживания, описанные в литературе. Среднее относительное отклонение составляет ~3.0%, медианное относительное отклонение составляет ~1.7%

для подмножеств NIST 17. Среднее абсолютное отклонение составляет ~34 единицы индекса удерживания. Рассмотрены только неполярные жидкие неподвижные фазы (такие как полидиметилсилоксан, 5% фенил 95% полидиметилсилоксан, сквалан). В ходе данной работы не делалось различия между различными видами неполярных жидких фаз. Ошибки, полученные с помощью разных методов машинного обучения и одинакового набора дескрипторов, сильно коррелируют между собой.

Ключевые слова: газовая хроматография, индекс удерживания, машинное обучение, градиентный бустинг.

Introduction

Gas chromatography is one of the most widely used methods of separation and chemical analysis. Hyphenated method gas chromatography – mass spectrometry is widely used for untargeted analysis, in particular for metabolomics and for the environmental analysis. The retention time highly depends on conditions of the chromatographic experiment. The retention index (RI) depends only on the chemical nature of a molecule and stationary phase. The retention index can be calculated based on the retention time [1].

Hence, reference RI can be used [2] for identification, for accepting or rejecting of candidate structures by comparison of observed and reference RI. Unfortunately, the reference retention index is available for less than 100.000 of chemical compounds in public databases [3]. It is several times less than a number of compounds for which the mass spectral information is available and almost thousand times less than a number of all known compounds.

So, the accurate and versatile RI prediction is an important task. The prediction task is the estimation of RI based on the structural formula of the compound. It can be realized by two approaches. The first approach is the physicochemical modeling of the chromatographic system using molecular mechanics [4]. Such works are mostly made for graphitized thermal carbon black. In recent years significant progress is achieved in this approach. The method can accurately predict retention for many classes of non-rigid organic molecules [5], for polychlorobiphenyls [6]. But such approach still is not universal and limited to selected classes of organic compounds, and the model has to be elaborated to support each new class. Also the method is limited to carbon stationary phases which are not the most widely used in practice.

The second approach is machine learning and quantitative structure-retention relationships [7]. Usually some molecular descriptors (MD) are calculated. MD are numerical values which characterize the molecule structure. MD are developed to be easy computable based on the structure and to be highly correlated with the chemical nature of the molecule. Different classes of MD are reviewed in [8]. An empirical function which allows to calculate RI from MD is developed for the RI prediction. Function parameters are selected to fit a training set of molecules with known RI. After parameters selection the relationship is validated using other set of molecules with known RI. Finally, after selection of all parameters and settings the relationship is tested using the third set. Usage of large and diverse sets for training (fitting of the function), validation and testing allows to achieve good accuracy and to ensure that the method really works for previously unseen molecules. String representation of a molecule structure converted to a matrix can be also used instead of conventional MD, but this approach requires the use of a convolutional neural network [9].

There are a lot of methods of machine learning regression, *i.e.,* methods to describe unknown complex real-world relationships such as relationship between MD and RI. Linear regression [10-12], support vector regression [12-14], artificial neural networks [15], *k*-nearest neighbors method [14], radial basis networks [12] are widely used for the RI prediction. Methods based on decision trees are not used widely.

Decision tree is a machine learning method which uses multiple decisions based on values of input variables for selection of a resulting value. Tree-like structure is processed

by the algorithm from root to leaves. Each node is a «question» based on input parameters. Depending on value of the parameters, one of the edges is selected during the processing. The final result is calculated based on a leaf on which the processing is finished. The tree and «tests» located in nodes are generated using a training set during the fitting.

Gradient boosting regression is a more complex method. It uses a few decision trees. Each tree corrects the results obtained after previous trees, *i.e.,* each tree, except the first one, fits the residuals between a true value and a result obtained using previous trees. The large number of trees with a limited number of nodes are generated sequentially. The detailed explanation of these algorithms can be found in [16-18].

Gradient boosting is extremely widely used and helpful machine learning method, but still there are no works on the use of gradient boosting for the RI prediction. The aim of this work is to examine if gradient boosting is efficient for the RI prediction using MD.

Methods

NIST 17 database was used for training, testing and validation [3]. All RI values for standard non-polar and semi-standard non-polar stationary phases were averaged together for every compound. Stereoisomers were treated as identical compounds. Three random subsets of NIST 17 RI database containing 42234, 5153 and 25589 molecules, were used for training, validation and testing respectively. A very few compounds which have RI > 5000 were excluded.

All calculations were provided with in-house software written using Java programming language. Smile (version 1.5.2) machine learning library [18] was used for gradient boosting. 8000 consequent trees with 6 leaves each were used. Shrinkage parameter and sampling fraction at each step were set at 0.03 and 0.7 respectively. The Huber loss function [17-18] was used. These hyperparameters were found using a grid search. The validation set was used for the hyperparameters search.

Two more RI data sets were used for testing: flavors and fragrances [10, 19] and essential oils [20]. Compounds from these sets are not contained in the training set. There is no overlapping between the training set and the sets used for testing and validation.

177 molecular descriptors calculated using Chemical Development Kit [21] were used. The set of descriptors and the scaling factors were exactly the same as used in [15].

A neural network with one hidden layer (90 hidden nodes) was used for the comparison. The neural network and learning procedure were as close to the ones used in the work [15] as possible. We used our own implementation based on Eclipse Deeplearning4j library [22]. Hyperbolic tangent and identity were employed as activation functions for input and output layers respectively. 50 epochs (full runs over the entire training data set) were performed (Adam optimization algorithm with learning rate 0.001). The mean absolute error was used as a loss function. The same data sets and the set of descriptors were used for the neural network and gradient boosting.

Results and discussion

Preliminary tuning of hyperparameters was performed using the validation set. Number of leaves was varied in range 2-10, number of trees in range 1000-10000, shrinkage parameter and sampling fraction in range 0-1. The grid search was used. For the neural network the hyperparameters from work [15] were used. We made an attempt to tune them, but neither changing a hidden nodes number or increasing a hidden layer number allow to significantly improve the performance of the neural network. The change of activation functions reduces accuracy.

Table 1 summarizes the performance of both methods for different data sets. The root mean square error (RMSE) is much higher than the mean absolute error (MAE) in all cases. MAE in its turn is much larger than the median absolute error (MdAE). A very few distant outliers cause it. The same situation is with the mean percentage error (MPE) and the median percentage error (MdPE).

Table 1. Accuracy of retention index prediction using gradient boosting and neural network with one hidden layer. Four data sets are considered.

	Gradient boosting					Neural network with one hidden layer				
Data set	MdAE	MdPE	MAE	MPE	RMSE	MdAE	MdPE	MAE	MPE	RMSE
Validation set	34.0	1.66	57.0	3.00	110.1	37.7	1.91	68.2	3.43	127.3
Test set	34.3	1.68	58.4	3.04	112.5	38.1	1.90	68.5	3.43	130.0
Flavors and fragrances	29.7	2.58	46.2	3.94	83.4	25.2	2.21	41.4	3.44	80.9
Essential oils	36.7	2.26	51.0	3.11	65.0	51.0	3.11	54.5	3.22	70.5

In all cases, except the flavors and fragrances data set, gradient boosting outperforms the neural network. The used neural network is state-of-the-art method of the RI prediction, which is used in METEXPERT expert system. Hence, it can be concluded that gradient boosting performs at the same or even better level than other modern RI prediction methods.

Achieved accuracy is higher than accuracy of other methods based on linear regression [10-11] and on support vector regression [13], but direct comparison is not possible because different data sets are used in different works. Gradient boosting shows almost the same accuracy comparing with the best at the moment RI prediction method – a deep convolutional neural network [9] based on string representation of a molecule. According to [9, 11, 13-14], even lesser precision is enough for improvement of a mass spectral library search and for rejecting of false candidates.

Figure 1 shows a correlation plot between predicted and reference RI and distribution of residuals for the test data set. The distribution of residuals can be approximated with the equation:

$$p(\Delta(RI)) = 0.2055\exp(-|\Delta(RI)|/49.06)$$

near zero and is highly tailed farther from zero. Correlation coefficient between predicted and reference RI is 0.97.

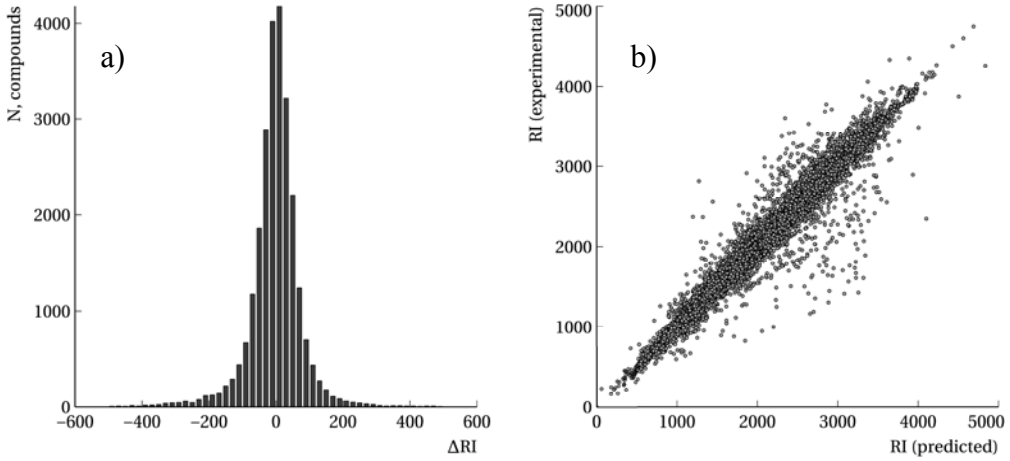

Fig. 1. Distribution of residuals (a) and correlation plot (b) for retention indices predicted using gradient boosting in comparison with reference values. Random test subset of NIST 17 database is used.

Figure 2 shows a correlation plot between predicted and reference RI for the flavors and fragrances data set. Results achieved using both methods are shown. The same compounds are outliers for both prediction methods. Probably for these compounds there are wrong experimental data or the selected set of MD does not represent these compounds well. In this work no custom MD selection was provided. More careful selection of MD, for example using genetic algorithm or use of more sophisticated MD probably will allow further improvement of the accuracy.

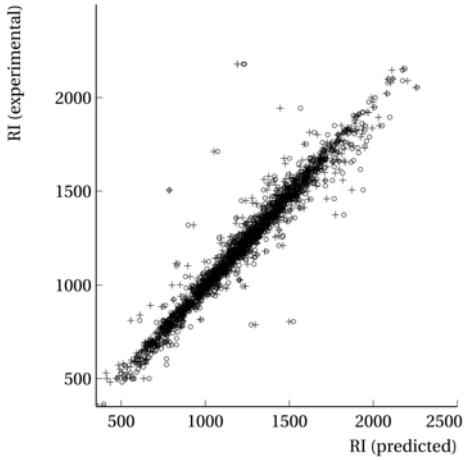

Fig. 2. Correlation plot between predicted and reference retention indices for flavors and fragrances data set. O – gradient boosting; + – neural network with one hidden layer.

In this work we also tested other decision trees based machine learning algorithms – single decision tree and random forest, but the achieved results were worse, maybe due to inappropriate hyperparameters tuning. Gradient boosting shows the best results of three decision trees based algorithms.

Conclusions

This work shows that gradient boosting can be used for the retention index prediction. Optimal values of hyperparameters are given. The precision obtained with this method in most cases is better than that obtained using a neural network with one hidden layer. Use of gradient boosting in quantitative retention-property relationships is still limited, but this work clearly shows the advantages of this method for the prediction of physicochemical characteristics of chemical compounds. Errors obtained with different machine learning algorithms and with the same representation of the molecule strongly correlate with each other. Non-polar liquid stationary phases are considered.

The work was supported by the Ministry of Science and Higher Education of the Russian Federation.

References

1. Zellner B.D.A., Bicchi C., Dugo P., Rubiolo P. et al., *Flavour Fragr. J.*, 2008, Vol. 23, No 5, pp. 297-314, DOI: 10.1002/ffj.1887

2. Zhang J., Koo I., Wang B., Gao Q.W. et al., *J. Chromatogr. A*, 2012, Vol. 1251, pp. 188-193, DOI: 10.1016/j.chroma.2012.06.036

3. Available at: https://chemdata.nist.gov/ (accessed 06 Nov 2019).

4. Buryak A.K., *Russ. Chem. Rev.*, 2002, Vol. 71, No 8, pp. 695-706, DOI: 10.1070/RC2002v071n08ABEH000711

5. Matyushin D.D., Buryak A.K., *Sorbtionnye I khromatograficheskie protsessy*, 2017, Vol. 17, No 2, pp. 204-211, DOI: 10.17308/sorpchrom.2017.17/372

6. Matyushin D.D., Buryak A.K., *J. Anal. Chem.*, 2019, Vol. 74, Supplement 1, pp. 47-51, DOI: 10.1134/S1061934819070165

7. Heberger K., *J. Chromatogr. A*, 2007, Vol. 1158, No 1-2, pp. 273-305, DOI: 10.1016/j.chroma.2007.03.108
8. Yap C.W., *J. Comput. Chem.*, 2011, Vol. 32, No 7, pp. 1466-1474, DOI: 10.1002/jcc.21707
9. Matyushin D.D., Sholokhova A.Yu., Buryak A.K., *J. Chromatogr. A*, 2019, Vol. 1607, pp. 460395, DOI: 10.1016/j.chroma.2019.460395
10. Rojas C., Duchowicz P.R., Tripaldi P., Diez R.P., *Chemom. Intell. Lab. Syst.*, 2015, Vol. 140, pp. 126-132, DOI: 10.1016/j.chemolab.2014.09.020
11. Kumari S., Stevens D., Kind T., Denkert C. et al., *Anal. Chem.*, 2011, Vol. 83, No 15, pp. 5895–5902, DOI: 10.1021/ac2006137
12. Chen H.F., *Anal. Chim. Acta*, 2008, Vol. 609, No 1, pp. 24-36, DOI: 10.1016/j.aca.2008.01.003
13. Mikhaleva V.V., Verhoeven H.A., De Vos R.C.H., van Ham R.C., *Bioinformatics*, 2009, Vol. 25, No 6, pp. 787-794, DOI: 10.1093/bioinformatics/btp056
14. Dossin E., Martin E., Diana P., Castellon A. et al., *Anal. Chem.*, 2016, Vol. 88, No. 15, pp. 7539–7547, DOI: 10.1021/acs.analchem.6b00868
15. Qiu F., Lei Z., Sumner L.W., *Anal. Chim. Acta*, 2018, Vol. 1037, pp. 316-326, DOI: 10.1016/j.aca.2018.03.052
16. Roe B.P., Yang H.-J., Zhu J., Liu Y. et al., *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2005, Vol. 543, No 2-3, pp. 577-584, DOI: 10.1016/j.nima.2004.12.018
17. Natekin A., Knoll A., *Frontiers in neurorobotics,* 2013, Vol. 7, pp. 21, DOI: 10.3389/fnbot.2013.00021
18. Available at: https://haifengl.github.io/ (accessed 28 Nov 2019).
19. Jennings W., Qualitative Analysis of Flavor and Fragrance Volatiles by Glass Capillary Gas Chromatography, London, Academic Press, INC, 1980, 472 p.
20. Adams R.P., Identification of Essential Oil Components by Gas Chromatography – Mass Spectrometry, 4th edition, USA, Allured publishing corporation, Carol Stream, 2007, Vol. 456, 804 p.
21. Willighagen E.L., Mayfield J.W., Alvarsson J., Berg A. et al., *J. Cheminformatics*, 2017, Vol. 9, No 1, p. 33, DOI: 10.1186/s13321-017-0220-4
22. Available at: http://deeplearning4j.org (accessed 06 Nov 2019).

Матюшин Дмитрий Дмитриевич – м.н.с. лаборатории физико-химических основ хроматографии и хромато-масс-спектрометрии, Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва

Шолохова Анастасия Юрьевна – м.н.с. лаборатории физико-химических основ хроматографии и хромато-масс-спектрометрии, Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва

Буряк Алексей Константинович – заведующий лабораторией физико-химических основ хроматографии и хромато-масс-спектрометрии, проф., д.х.н. Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва

Matyushin Dmitriy Dmitrievich – junior researcher, laboratory of physicochemical principles of chromatography and chromatography – mass spectrometry; Institute of Physical Chemistry and Electrochemistry, Moscow, e-mail: dm.matiushin@mail.ru

Sholokhova Anastasia Yur'evna – junior researcher, laboratory of physicochemical principles of chromatography and chromatography – mass spectrometry; Institute of Physical Chemistry and Electrochemistry, Moscow, e-mail: shonastya@yandex.ru

Buryak Aleksey Konstantinovich – prof., grand PhD (chemistry), laboratory of physicochemical principles of chromatography and chromatography – mass spectrometry; Institute of Physical Chemistry and Electrochemistry, Moscow, e-mail: akburyak@mail.ru