

А. Д. Лифанов, К. Ю. Никитина, Е. Г. Лифанова

ПРОГНОЗИРОВАНИЕ ОКТАНОВЫХ ЧИСЕЛ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Ключевые слова: большие данные, индустрия 4.0, октановое число, машинное обучение, искусственный интеллект.

В настоящее время нефтехимические процессы направлены на получение бензиновых фракций с высоким октановым числом. Октановое число – это показатель, который характеризует устойчивость топлива к детонации (взрыву) в двигателях внутреннего сгорания. Чем выше данный показатель, тем большее количество энергии оно способно производить. В качестве эталона был взят изооктан (2,2,4-триметилпентан). Иногда используют его смесь с нормальным гептаном. Для повышения октановых чисел бензиновых фракций обычно вводят антидетонационные присадки. В качестве антидетонационных присадок используют органические соединения различного строения. Обычно это изоалканы, олефины, циклоалканы, спирты и т.д. с высокими октановыми числами. Особенности влияния антидетонационных присадок на октановые числа топлив посвящены многочисленные работы. В данных работах показана возможность направленного регулирования октановых чисел бензиновых фракций путем варьирования природы модифицирующих добавок и их количества. Обычно наблюдается аддитивное увеличение октанового числа бензиновых фракций пропорционально содержанию антидетонационной присадки. Однако возможны и исключения, которые устанавливаются эмпирически путем. Наиболее перспективным является установление закономерностей связи химического строения углеводородов с показателями детонации топлив. В данном подходе органические соединения представляется в виде набора дескрипторов, характеризующих особенности химической структуры молекул. Данные об октановых числах для органических соединений были взяты из базы данных PubChem. В базу данных для данной работы была включена информация об октановых числах для 219 органических веществ. Для упрощения анализа представлений органических соединений, мы использовали ключи MACC, поскольку они являются одними из лучших дескрипторов для прогнозирования свойств химических соединений. Данные дескрипторы создаются на основе общих ключей подструктуры. Кроме того, модели были рассчитаны с использованием молекулярных отпечатков Морган, также известных как циркулярные отпечатки с радиусом 2. В рамках данной работы были реализованы гребневая регрессия, алгоритм случайного леса, метод ближайших соседей kNN, метод опорных векторов (SVM), искусственные нейронные сети. Для обучающей выборки полученная классификационная модель случайного леса показала безошибочную классификацию, ошибка прогноза для нее равна 0. Статистические характеристики построенной XGBoost-модели для выборки имеют следующие значения: $R^2 = 0.83$ и ошибкой предсказания $RMSE = 9.96$ °C.

A. D. Lifanov, K. Yu. Nikitina, E. G. Lifanova

PREDICTION OF OCTANE NUMBERS OF ORGANIC COMPOUNDS USING MACHINE LEARNING METHODS

Keywords: big data, industry 4.0, octane number, machine learning, artificial intelligence.

Currently, petrochemical processes are aimed at producing gasoline fractions with a high octane number. The octane number is an indicator that characterizes the resistance of fuel to detonation (explosion) in internal combustion engines. The higher this indicator is the more energy it is able to produce. Isooctane (2,2,4-trimethylpentane) was used as a reference. Sometimes a mixture of it with normal heptane is used. Anti-knock additives are usually introduced to increase the octane numbers of gasoline fractions. Organic compounds of various structures are used as anti-knock additives. These are usually isoalkanes, olefins, cycloalkanes, alcohols, etc. with high octane numbers. Numerous works have been devoted to the peculiarities of the effect of anti-knock additives on the octane numbers of fuels [1-8]. These studies show the possibility of directional control of the octane numbers of gasoline fractions by varying the nature of modifying additives and their amount. An additive increase in the octane number of gasoline fractions is usually observed in proportion to the content of the anti-knock additive. However, there may be exceptions that are established empirically. The most promising is the establishment of patterns of connection between the chemical structure of hydrocarbons and the detonation of fuels. In this approach, organic compounds are represented as a set of descriptors characterizing the features of the chemical structure of molecules. Flash point data for organic compounds were taken from the PubChem database. Information on the flash point for 219 organic substances was included in the database for this work. To simplify the analysis of the representation of organic compounds, we used MACC keys, as they are among the best descriptors for predicting the properties of chemical compounds. These descriptors are created based on the shared keys of the substructure. In addition, the models were calculated using Morgan molecular prints, also known as circular prints with a radius of 2. Within the framework of this work, ridge regression, the random forest algorithm, the kNN nearest neighbor method, the support vector machine (SVM) method, and artificial neural networks were implemented. For the training sample, the resulting classification model of a random forest showed an error-free classification, the prediction error for it is 0. The statistical characteristics of the constructed XGBoost model for the sample have the following values: $R^2 = 0.83$ and $RMSE = 9.96$ °C.

Введение

В настоящее время нефтехимические процессы направлены на получение бензиновых фракций с вы-

соким октановым числом. Октановое число – это показатель, который характеризует устойчивость топлива к детонации (взрыву) в двигателях внутреннего сгорания. Чем выше данный показатель, тем большее

количество энергии оно способно производить. В качестве эталона был взят изооктан (2,2,4-триметилпентан). Иногда используют его смесь с нормальным гептаном. Для повышения октановых чисел бензиновых фракций обычно вводят антидетонационные присадки.

В качестве антидетонационных присадок используют органические соединения различного строения. Обычно это изоалканы, олефины, циклоалканы, спирты и т.д. с высокими октановыми числами. Особенности влияния антидетонационных присадок на октановые числа топлив посвящены многочисленным работы [1-3]. В данных работах показана возможность направленного регулирования октановых чисел бензиновых фракций путем варьирования природы модифицирующих добавок и их количества. Обычно наблюдается аддитивное увеличение октанового числа бензиновых фракций пропорционально содержанию антидетонационной присадки. Однако возможны и исключения, которые устанавливаются эмпирически путем.

Наиболее перспективным является установление закономерностей связи химического строения углеводородов с показателями детонации топлив. В данном подходе органические соединения представляется в виде набора дескрипторов, характеризующих особенности химической структуры молекул.

К сожалению, для построения моделей «структура-свойство» для прогнозирования октановых чисел органических соединений отсутствует информация в крупных базах данных, например, PubChem [4]. Поэтому нами вручную собран набор данных, содержащий информацию об исследовательских (RON, англ. research octane number) и моторных (MON, англ. motor octane number) октановых числах.

Цель работы заключалась в исследовании закономерностей связи химического строения органических соединений с октановыми числами.

Материалы и методы

Традиционно, построение QSPR-модели проходит несколько этапов: подготовка данных, выбор модели, отбор дескрипторов, обучение модели и анализ результатов.

Подготовка данных. Данные об октановых числах для органических соединений были взяты из базы данных PubChem [5].

Собранный набор данных содержал информацию о свойствах 219 органических соединений с их значениями октановых чисел в диапазоне от -3 до 120.15°C.

В нашей работе в качестве линейных представлений выбрана система SMILES [6, 7]. Далее на основе SMILES генерируется дескриптор, отражающий особенности структуры органического соединения [8]. С помощью дескрипторов кодируется информация о структуре химического соединения. Описание структуры химических соединений в данном подходе может осуществляться по-разному, то есть разными наборами дескрипторов, выбор которых определяется степенью детализации структуры в решаемой задаче [9-13].

Данные дескрипторы рассчитаны с использованием библиотеки RDKit 2 [9]. RDKit представляет собой инструмент для хемоинформатики с открытым исходным кодом [9].

В качестве дескрипторов, отражающих структуры органических соединений, – молекулярные отпечатки (Morgan's fingerprints) [13] с параметрами: радиус 2, длина битовой строки 2048 бит (чем больше радиус, тем большее окружение атомов учитывается при кодировании фрагментов; чем больше длина битовой строки, тем больше информации о структурных фрагментах можно закодировать и избежать потери информации о связях «структура-свойство»), а также структурные ключи (MACCS keys) [10, 11].

Полученные молекулярные отпечатки и структурные ключи используются в качестве входных данных для обучения модели, которая может быть реализована на основе любого метода машинного обучения.

В рамках данной работы были реализованы гребневая регрессия, алгоритм случайного леса [14], метод ближайших соседей kNN [15, 16], метод опорных векторов (SVM) [17], градиентный бустинг (XGBoost) [18], AdaBoostRegressor (AdaBoost) [19].

Важным этапом является деление данных на тренировочные и тестовые [20]. В нашей модели в качестве тестовых данных взято 20% данных с помощью встроенной функции Scikit-learn - Train test split, которая случайным образом отбирает заданное количество данных для дальнейшей проверки [21].

Для оценки надежности моделей «структура-свойство» большое значение имеет валидация моделей с использованием внешней тестовой выборки, состоящей из соединений, которые не вошли в обучающую выборку. Если выборка соединений небольшая, то может проводиться процедура n-кратного скользящего контроля (n-folds).

Метод n-кратного скользящего контроля k-fold используется для оценки эффективности и возможности обобщения модели. Он заключается в разделении набора данных на k подмножеств (блоков) и итеративном использовании k-1 блоков для обучения модели и одного блока для перекрестной проверки модели. Процесс повторяется k раз, где каждый блок используется для перекрестной проверки ровно один раз.

Процедура обучения модели может включать определение гиперпараметров модели и их оптимизацию. Данная процедура осуществлялась с помощью инструмента GridSearchCV. В алгоритме случайного леса в качестве гиперпараметров выбирают число деревьев, глубину построения дерева и т.д. В случае метода ближайших соседей, гиперпараметрами модели являются число соседей k.

Для прогнозирования выбирается размер подмножества горизонтом в 1/5. Первоначально тренировочные данные используются для прогнозирования первого подмножества с 20% обучающих данных, затем данное подмножество включается в обучающие данные для прогнозирования следующего подмножества, и таким образом модель повторно обучается и обновляет веса параметров.

Метод кросс-валидации k-fold предоставляет более надежную оценку эффективности модели по

сравнению с одноразовым разделением на обучающий и тестовый наборы и позволяет избежать переобучения модели.

Коэффициент детерминации (R^2) рассчитывается следующим образом:

$$R^2 = 1 - \frac{\sum_i (y_i^{\text{pred}} - y_i^{\text{exp}})^2}{\sum_i (y_i^{\text{pred}} - \bar{y}^{\text{exp}})^2} \quad (1)$$

где y_i^{pred} и y_i^{exp} - предсказанные и истинные значения характеристик объектов, соответственно, N - объем выборки.

Как и коэффициент детерминации R^2 (1), статистический показатель Q^2 (тот же R^2 , после процедуры кросс-валидации) не может принимать значения, превышающие 1, причем, чем ближе его значение к 1, тем выше средняя прогнозирующая способность моделей [22].

Следующим важным статистическим критерием является среднеквадратичная ошибка (RMSE) (2). Среднеквадратичная ошибка RMSE может принимать любое положительное значение, включая ноль, причем, чем ниже, тем выше средняя прогнозирующая способность моделей, оцененная посредством процедуры 5-кратного скользящего контроля. RMSE рассчитывается следующим образом:

$$RMSE = \sqrt{\frac{\sum_i (y_i^{\text{pred}} - y_i^{\text{exp}})^2}{N - 1}} \quad (2)$$

Таким образом, показатели RMSE и Q^2 являются универсальными показателями прогнозирующей способности моделей [21]. Показатели RMSE и Q^2 (тот же R^2 , но с процедурой n -кратного перекрестного скользящего контроля) являются универсальными показателями прогнозирующей способности моделей [11].

Результаты и их обсуждение

В результате исследования мы получили адекватные модели «структура-свойство» с высокими статистическими показателями. Для прогноза и интерпретации использовались модели как результат пятикратного скользящего контроля (Рис. 1) [21].

В результате моделирования мы построили адекватную модель гребневой регрессии с использованием молекулярных отпечатков Моргана в качестве дескрипторов с коэффициентом детерминации для тестовой выборки $R^2 = 0.83$ и ошибкой предсказания $RMSE = 9.96^\circ\text{C}$.

В результате анализа фрагментов углеводородов не выявили четкой тенденции между вкладами фрагментов. Таким образом, мы получили адекватные регрессионные модели, на основе которых проведен прогноз определенных классов соединений для соответствующих свойств, а также провели сравнительный анализ фрагментов, способствующих или препятствующих их детонации в двигателях внутреннего сгорания.

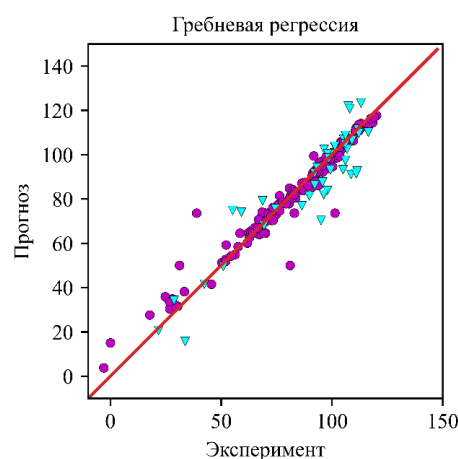


Рис. 1 – Регрессионная модель гребневой регрессии для прогнозирования октановых чисел органических соединений

Fig. 1 – A ridge regression model for predicting the octane numbers of organic compounds

Таблица 1 - Результаты прогнозирования октановых чисел органических соединений (RON) с использованием регрессионных моделей

Table 1 - The results of forecasting the octane numbers of organic compounds (RON) using regression models

Модель	Дескрипторы	Без кросс-валидации		С кросс-валидацией	
		R^2	RMSE	R^2	RMSE
Ridge	MACCS	0,52	16,88	0,52	16,88
	ECPF4	0,83	9,96	0,83	9,96
AdaBoost	MACCS	0,42	18,54	0,33	19,88
	ECPF4	0,67	13,97	0,64	14,66
RF	ECPF4	0,80	10,88	0,47	17,68
	MACCS	0,45	18,13	0,43	18,41
XGBoost	MACCS	0,54	16,49	0,45	18,13
	ECPF4	0,78	11,38	0,79	11,29
kNN	MACCS	0,63	14,76	0,67	14,08
	ECPF4	0,66	14,23	0,68	13,74
SVR	MACCS	0,33	19,90	0,48	17,64
	ECPF4	0,11	22,94	0,80	10,77

Таблица 2 - Результаты прогнозирования октановых чисел органических соединений (MON) с использованием регрессионных моделей

Table 2 - The results of forecasting the octane numbers of organic compounds (MON) using regression models

Модель	Дескрипторы	Без кросс-валидации		С кросс-валидацией	
		R^2	RMSE	R^2	RMSE
Ridge	MACCS	0,48	14,02	0,48	14,02
	ECPF4	0,86	7,35	0,86	7,35
AdaBoost	MACCS	0,30	16,15	0,28	16,39
	ECPF4	0,61	12,07	0,71	10,36
RF	ECPF4	0,85	7,39	0,58	12,56
	MACCS	0,35	15,64	0,34	15,71
XGBoost	MACCS	0,48	13,94	0,31	16,04
	ECPF4	0,82	8,25	0,85	7,61
kNN	MACCS	0,55	12,95	0,54	13,13
	ECPF4	0,68	10,97	0,79	8,83
SVR	MACCS	0,30	16,22	0,41	14,93
	ECPF4	0,14	17,94	0,86	7,36

Работа выполнена в рамках реализации Программы развития ФГБОУ ВО «КНИТУ» (ПСАЛ Приоритет 2030).

Литература

1. S. Fei, Y. Qi, W. Liu, Y. Wang, Z. Wang, H. Zhang, *Combustion Science and Technology*, **7**, 1-19 (2021).
2. K. Oyama, T. Sone, *Journal of the Japan Petroleum Institute*, **54**, 3, 201-207 (2011).
3. М.Ю. Доломатов, О.С. Коледин, Э.А. Ковалева, Р.А. Федина, Р.В. Гарипов, М.Р. Валеев, *Башкирский химический журнал*, **29**, 2, 65-70 (2022).
4. T.I. Madzhidov, A. Rakhimbekova, V.A. Afonina, *Mendeleev Communications*, **31**, 6, 769-780 (2021).
5. <http://www.chemspider.com/>.
6. D. Weininger, *Journal of Chemical Information and Computer Sciences*, **28**, 1, 31-36 (1988).
7. D. Weininger, A. Weininger, J.L. Weininger, *Journal of Chemical Information and Computer Sciences*, **29**, 2, 97-101 (1989).
8. R. Todeschini, V. Consonni, New York: Wiley-VCH, 1-680 (2000).
9. <https://rdkit.org/>
10. J. Gasteiger, T. Engel, Weinheim: Wiley-VCH (2003).
11. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, *J. Chem. Inf. Comput. Sci.*, **42**, 6, 1273-1280 (2002).
12. О.С. Коледин, М.Ю. Доломатов, Р.Ш. Япаев, А.Т. Гильмутдинов, М.Ф. Мухарметов, Р.В. Гарипов, М.Р. Валеев, *Журнал прикладной химии*, **95**, 5, 666-671 (2022).
13. H.L. Morgan, *Journal of Chemical Documentation*, **5**, 2, 107-113 (1965).
14. L. Breiman, *Machine Learning*, **45**, 1, 5-32 (2001).
15. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Consistency properties. Technical report. USAF School of Aviation Medicine (1951).
16. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Small sample performance. Technical report. USAF School of Aviation Medicine (1952).
17. Cortes, C., V. Vapnik, Support-vector networks. *Machine Learning*, **20**, 3, 273-297 (1995).
18. S.B. Gunturi, K. Archana, A. Khandelwal, R. Narayanan, *QSAR & Combinatorial Science*, **27**, 11-12, 1305-1317 (2008).
19. M. Budka, B. Gabrys, *Procedia Computer Science*, **1**, 1, 193-201 (2010).
20. Schein, A.I., L.H. Ungar, *Machine Learning*, **68**, 3, 235-265 (2007).
21. Gramatica, P. *QSAR & Comb Sci*, 694-701 (2007).
22. Meloun M, Militku J, Hill M. *Analyst*, **127**, 433-450 (2002).
23. Tharwat A. *Applied Computing and Informatics*, **17**, 1, 168-192 (2021).

References

1. S. Fei, Y. Qi, W. Liu, Y. Wang, Z. Wang, H. Zhang, *Combustion Science and Technology*, **7**, 1-19 (2021).
2. K. Oyama, T. Sone, *Journal of the Japan Petroleum Institute*, **54**, 3, 201-207 (2011).
3. M.Y. Dolomatov, O.S. Koledin, E.A. Kovaleva, R.A. Fedina, R.V. Garipov, M.R. Valeev, *Bashkir Chemical Journal*, **29**, 2, 65-70 (2022).
4. T.I. Madzhidov, A. Rakhimbekova, V.A. Afonina, *Mendeleev Communications*, **31**, 6, 769-780 (2021).
5. <http://www.chemspider.com/>.
6. D. Weininger, *Journal of Chemical Information and Computer Sciences*, **28**, 1, 31-36 (1988).
7. D. Weininger, A. Weininger, J.L. Weininger, *Journal of Chemical Information and Computer Sciences*, **29**, 2, 97-101 (1989).
8. R. Todeschini, V. Consonni, New York: Wiley-VCH, 1-680 (2000).
9. <https://rdkit.org/>
10. J. Gasteiger, T. Engel, Weinheim: Wiley-VCH (2003).
11. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, *J. Chem. Inf. Comput. Sci.*, **42**, 6, 1273-1280 (2002).
12. O.S. Koledin, M.Yu. Dolomatov, R.Sh. Yapaev, A.T. Gil'mutdinov, M.F. Muharmetov, R.V. Garipov, M.R. Valeev, *Russian Journal of Applied Chemistry*, **95**, 5, 666-671 (2022).
13. H.L. Morgan, *Journal of Chemical Documentation*, **5**, 2, 107-113 (1965).
14. L. Breiman, *Machine Learning*, **45**, 1, 5-32 (2001).
15. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Consistency properties. Technical report. USAF School of Aviation Medicine (1951).
16. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Small sample performance. Technical report. USAF School of Aviation Medicine (1952).
17. Cortes, C., V. Vapnik, Support-vector networks. *Machine Learning*, **20**, 3, 273-297 (1995).
18. S.B. Gunturi, K. Archana, A. Khandelwal, R. Narayanan, *QSAR & Combinatorial Science*, **27**, 11-12, 1305-1317 (2008).
19. M. Budka, B. Gabrys, *Procedia Computer Science*, **1**, 1, 193-201 (2010).
20. Schein, A.I., L.H. Ungar, *Machine Learning*, **68**, 3, 235-265 (2007).
21. Gramatica, P. *QSAR & Comb Sci*, 694-701 (2007).
22. Meloun M, Militku J, Hill M. *Analyst*, **127**, 433-450 (2002).
23. Tharwat A. *Applied Computing and Informatics*, **17**, 1, 168-192 (2021).

© А. Д. Лифанов – к.х.н., доцент кафедры Общей химической технологии, Казанский национальный исследовательский технологический университет (КНИТУ), Казань, Россия, lifanov84@mail.ru; К. Ю. Никитина – студент кафедры Химии и технологии переработки эластомеров, КНИТУ, karina2004nikitina@mail.ru; Е.Г. Лифанова – ассистент кафедры Дизайна и национальных искусств, Казанский (Приволжский) федеральный университет, Казань, Россия, 678769@mail.ru.

© A. D. Lifanov – PhD (Chemical Sci.), Associate Professor of the Department of General Chemical Technology, Kazan National Research Technological University (KNRTU), Kazan, Russia, lifanov84@mail.ru; K. Yu. Nikitina – Student of the Department of Chemistry and Processing Technology of Elastomers, KNRTU, karina2004nikitina@mail.ru; E. G. Lifanova – Assistant of the Department of Design and National Arts, Kazan (Volga Region) Federal University, Kazan, Russia, 678769@mail.ru.