UDC 004.94 *Научная статья*

https://doi.org/10.23947/2587-8999-2023-6-1-77-82

Predicting the kinetics of complex luminescence processes in Python

S. V. Slepnev[1], K. F. Koledina[1,2] ✉

[1]Ufa State Petroleum Technical University, 1, Kosmonavtov St., Ufa, Russian Federation

[2]Institute of Petrochemistry and Catalysis of the Russian Academy of Sciences, 141, October Ave, Ufa, Russian Federation

✉ koledinakamila@mail.ru

Abstract

Introduction. Polyarylene phthalides (PAF) are widely used in optoelectronics today. The reactions occurring during the synthesis of polyarylene phthalides have a complex character, which has not yet been described using mathematical models. In this regard, it is impossible to use PAF in many processes. Polyarylene phthalides have luminescence, good optical and electrophysical properties. The elucidation of the mechanisms of the occurrence of luminescent states of PAF is of both fundamental and practical interest.

The elucidation of the mechanisms of the occurrence of luminescent states of PAF is of both fundamental and practical interest. Due to the complexity of calculating the kinetics of the luminescence intensity of polyarylene phthalides using known mathematical models, the aim of the study was to build a system using machine learning methods that predicts luminescence values depending on temperature and heating time.

Materials and methods. Experimental data have been prepared for calculations, the use of "random forest" and "gradient boosting" methods has been justified, a method for selecting hyperparameters of these models has been selected and the expediency of its use has been justified, optimal models have been constructed and predictions have been obtained.

The results of the study. An algorithm for predicting the luminescence intensity of polyarylene phthalides has been developed. Using machine learning methods based on experimental data, the key hyperparameters of the system were determined and the average accuracy of predicting values was achieved — 80 %.

Discussion and conclusions. High-accuracy forecasts will allow predicting how products containing polyarylene phthalides will react to external influences. The paper presents two methods for solving the problem, as they showed the best results.

Keywords: Python, luminescence, machine learning methods, random forest, gradient boosting.

For citation. Slepnev, S. V. Forecasting the kinetics of complex luminescence processes in Python / S. V. Slepnev, K. F. Koledina // Computational Mathematics and Information Technologies. — 2023. — Vol. 6, no. 1. — P. 77–82. https://doi.org/10.23947/2587-8999-2023-6-1-77-82

Original article

Прогнозирование кинетики сложных процессов люминесценции на Python

С. В. Слепнёв[1], К. Ф. Коледина[1,2] ✉

[1]Уфимский государственный нефтяной технический университет, Российская Федерация, г. Уфа, ул. Космонавтов, 1

[2]Институт нефтехимии и катализа УФИЦ РАН, Российская Федерация, г. Уфа, пр. Октября, 141

✉ koledinakamila@mail.ru

Аннотация

Введение. На сегодняшний день полиариленфталиды (ПАФ) находят широ-кое применение в оптоэлектронике. При этом реакции, протекающие при синтезе полиариленфталидов, имеют сложный характер, который до сих пор не удалось описать с помощью математических моделей. В связи с этим, невозможно использовать ПАФ во многих процессах. При этом ПАФ обладают люминесценцией, хорошими оптическими и электрофизическими свойствами. Выяснение механизмов возникновения люминесцирующих состояний ПАФ представляет как фундаментальный, так и практический интерес. Выяснение механизмов возникновения люминесцирующих состояний ПАФ представляет как фундаментальный, так и практический интерес. В связи со сложностью расчета кинетики интенсивности свечения полиариленфталидов с помощью известных математических моделей была поставлена цель исследования — построить с помощью методов машинного обучения систему, прогнозирующую значения люминесценции в зависимости от температуры и времени нагревания.

Материалы и методы. Подготовлены к вычислениям экспериментальные данные, обосновано использование методов «случайный лес» и «градиентный бустинг», выбран способ подбора гиперпараметров данных моделей и обоснована целесообразность его использования, построены оптимальные модели и получены предсказания.

Результаты исследования**.** Разработан алгоритм предсказания интенсивности свечения полиариленфталидов. Используя методы машинного обучения на экспериментальных данных, были определены ключевые гиперпараметры системы и достигнута средняя точность предсказания значений — 80 %.

Обсуждение и заключения. Прогнозы высокой точности позволят предсказывать, как будут реагировать на внешнее воздействие продукты, включающие в свой состав полиариленфталиды. В работе представлено два метода решения задачи, так как они показали наилучшие результаты.

Ключевые слова: Python, люминесценция, методы машинного обучения, случайный лес, градиентный бустинг.

Для цитирования. Слепнёв, С. В. Прогнозирование кинетики сложных процессов люминесценции на Python / С. В. Слепнёв, К. Ф. Коледина // Computational Mathematics and Information Technologies. — 2023. — Т. 6, № 1. — С. 77–82. https://doi.org/10.23947/2587-8999-2023-6-1-77-82

Introduction. Organic polymer materials are widely used in optoelectronics today. One of the varieties of polymers suitable for these purposes may be polyarylene phthalides (PAF). PAF are characterized by high thermal and chemical resistance, high film-forming properties. PAF have luminescence, good optical and electrophysical properties. The elucidation of the mechanisms of the occurrence of luminescent states of PAF is of both fundamental and practical interest. It is assumed that the luminescence of PAF is due to the formation of active intermediates under energetic action on the polymer, but their chemical nature and properties have not been studied at all [1].

Since the reactions occurring during the synthesis of polyarylene phthalides are complex, the need for detailed research in this practically unexplored area is quite obvious. However, despite the research and experiments carried out [1, 2, 3], a mathematical model describing the behavior of the glow of polyarylene phthalides has not been built.

Due to the complexity of calculating the kinetics of the luminescence intensity of polyarylene phthalides using well–known mathematical models [2, 3], the goal was set to build a system using machine learning methods that predicts luminescence values depending on temperature and heating time.

The following tasks were set to achieve the goal:

– to prepare experimental data for calculations;

– to analyze machine learning algorithms and create a program using the most optimal methods;

– to choose the most successful hyperparameters for models.

Materials and methods. Polyarylene phthalides are a type of aromatic polymers. It is assumed that the luminescence of polyarylene phthalides is due to the formation of active intermediates under energetic action on the polymer. In experimental studies, the heating temperature of the polyarylene phthalide membrane was changed in the range

from 298 to 460 K for several hours, using different rates of controlled heating and cooling of the polyarylene phthalide membrane. They found that the temperature effect on the membrane leads to the appearance of a long-fading glow of recombination luminescence [1].

PAF membrane was irradiated with unfiltered light for 10 minutes (100 W lamp), after which it was left at a temperature of 298 K for 8 hours to accumulate stable radical ions. The temperature of the PAF film was changed in the range from 298 to 460 K for several hours, using different speeds of controlled heating and cooling of the PAF membrane.

Data on the characteristics of individual polyarylene phthalides are presented in the following categories: time, temperature, luminescence intensity. There are 20 experiments in total, 200 values for each characteristic.

Various machine learning methods were tested on the obtained data and the most suitable models were selected for subsequent improvement. The selection of hyperparameters is implemented through a randomized search for parameters, each of which is selected from a distribution of possible values.

Random forest. The Random Forest algorithm is a universal machine learning algorithm, the essence of which is to use an ensemble of decision trees. The decision tree itself provides an extremely low quality of classification, but due to the large number of them, the result is significantly improved [4].

Compared to other machine learning methods, the theoretical part of the Random Forest algorithm is simple. There is no large amount of theory, only the formula of the final classifier is needed $a(x)$:

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x), \tag{1}$$

where N is the number of trees; i is the counter for trees; b is the decision tree; $a(x)$ is the sample generated from the data.

However, despite its universality, this method has a number of significant drawbacks [5]:

– complexity of interpretation;

– random forest can't extrapolate;

– the algorithm is prone to retraining on highly noisy data;

– for data involving categorical variables with a different number of levels, random forests are biased in favor of features with a large number of levels;

– larger size of the resulting models. Requires $O(N \cdot C)$ memory to store the model, where C is the number of trees.

Gradient boosting. It is a machine learning method that creates a crucial prediction model in the form of an ensemble of weak prediction models, usually decision trees. He builds the model in stages, allowing to optimize an arbitrary differentiable loss function [6].

Let L be a differentiable loss function, and algorithm $a(x)$ is a composition of basic algorithms:

$$a(x) = a_k(x) = b_1(x) + ... + b_k(x), \tag{2}$$

where the basic algorithm b_k is trained so as to improve the predictions of the current composition:

$$b_k = \arg\min \sum_{i=1}^{N} L(y_i, a_{k-1}(x_i) + b(x_i)). \tag{3}$$

The b_0 model is chosen in such a way as to minimize losses on the training sample:

$$b_0 = \arg\min \sum_{i=1}^{N} L(y_i, b(x_i)). \tag{4}$$

Consider the Taylor decomposition of the loss function L to build basic algorithms in the following steps up to the first term in the neighborhood of the point$(y_i, a_{k-1}(x_i))$:

$$L(y_i, a_{k-1}(x_i)) := L(y_i, a_{k-1}(x_i)) + b(x_i)\, g_i^{k-1}. \tag{5}$$

Let's get an optimization problem by getting rid of the permanent members:

$$b_k \approx \arg\min \sum_{i=1}^{N} b(x_i) g_i^{k-1}. \tag{6}$$

The basic algorithms b_k are trained to predict the values of the anti-gradient of the loss function from the current predictions of the composition at each iteration.

The main disadvantages of this method include the susceptibility to retraining and the voting system of appraisers. The vote of appraisers in gradient boosting is unequal. Some appraisers have a higher weight than others. As a rule, the vote of the very first trained appraiser has the lowest weight, and the last appraiser has the highest weight when voting.

Research results. The regression problem was solved using the Python programming language, and the accuracy of the solution was estimated through the coefficient of determination, which is calculated by the formula:

$$R^2 = 1 - \frac{S_{res}}{S_{tot}}, \tag{7}$$

where S_{res} is the sum of squares of residual errors, and S_{tot} is the total sum of errors.

Fig. 1, 2 shows the results of modeling the kinetics of the intensity of the PAF glow depending on the temperature change. The red curve shows experimental data on the intensity of the glow after exposure to temperature (the blue dotted curve along the auxiliary vertical axis).

Figure 1 shows intensity modeling by gradient boosting (green curve). Figure 2 shows the intensity simulation by the random forest method (green curve).

The coefficient of determination, when using gradient boosting, is 83 %, and in the case of using a random forest — 88 %.

The results turned out to be slightly different from each other due to the difference in methods, as expected. When using the gradient boosting method, the prediction turns out to be less accurate, and peaks are predicted at those times when they did not exist. In the case of a random forest, there are predictions that lie closer to the experimental data. Since gradient boosting copes with temperature prediction worse, this affects the predictive ability of the entire model. Such a result can be justified by an imperfect selection of hyperparameters and an abrupt change in the intensity of the glow, which makes it difficult to determine the direction of the gradient of the function.

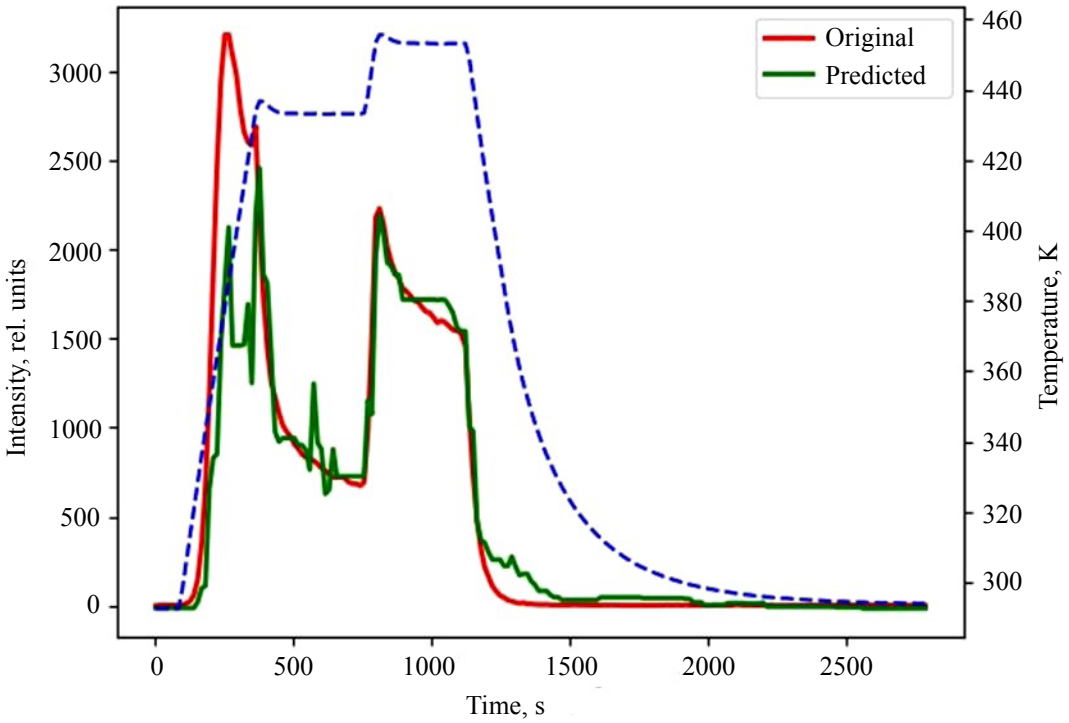

Fig. 1. Modeling of the kinetics of the intensity of the PAF glow by the Gradient boosting method

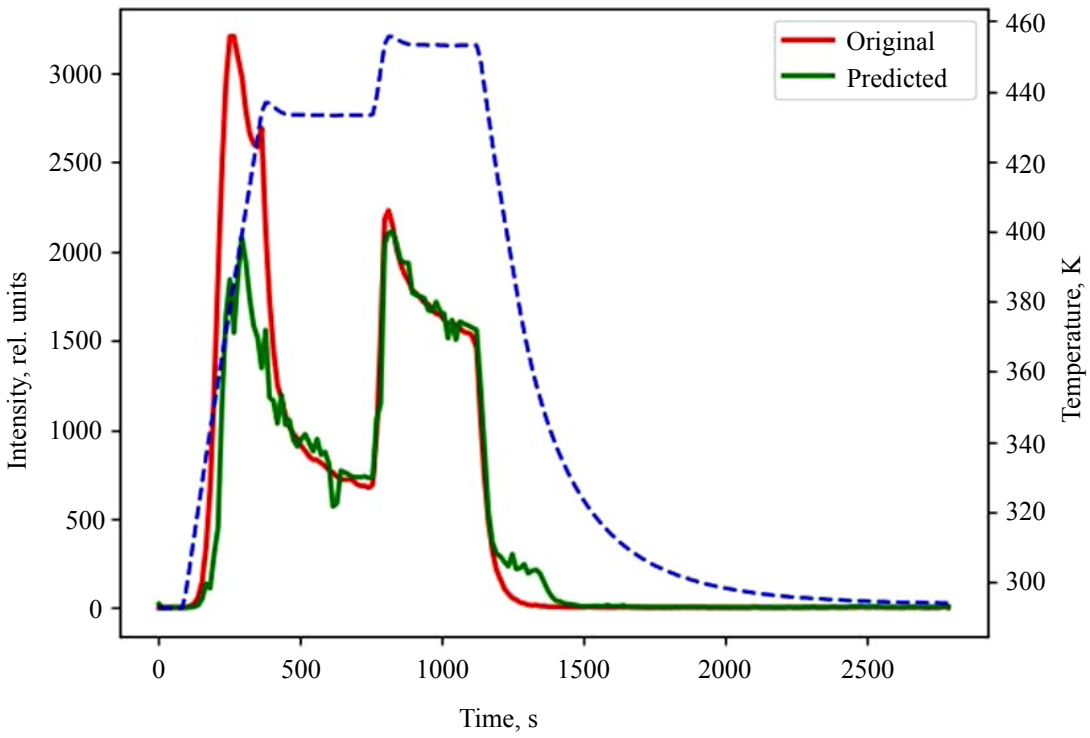

Fig. 2. Modeling of the kinetics of the intensity of the PAF glow by the Random forest method

Discussion and conclusions. The result obtained in the course of solving the problem of predicting the luminescence intensity of polyarylene phthalides is relevant for use in industrial and laboratory processes. Despite the high predictive power of this model, it tends to often make mistakes when predicting peaks, which requires further refinement. To reduce the modeling error, it is possible to resort to combining approaches and using other methods to obtain greater accuracy. An additional way to improve the model is also a more accurate selection of hyperparameters. It, in turn, requires high computing power, since it uses a search of all possible combinations and can be carried out for weeks for the current task. A potential solution to the problem may be the preliminary analysis of data using genetic algorithms to search for extremes and the subsequent transfer of data for training machine learning algorithms.

References

1. Ovchinnikova, M. Yu. Kinetic model of thermally stimulated luminescence of polydiphenylene phthalide films / M. Yu. Ovchinnikova,V. A. Antipin, S. L. Khursan // Kinetics and catalysis. — 2019. — Vol. 60, no. 5. — P. 547–554.

2. Koledina, K. F. Solving the problem of multi-criteria optimization of the synthesis reaction of benzylalkyl esters by the method of “ideal” point and lexicographic ordering / K. F. Koledina, A. A. Alexandrova // Computational mathematics and information technologies. — 2022. — Vol. 1, no. 1. — P. 12–19 Doi 10.23947/2587-8999-2022-1-1-12-19

3. Shaimardanova, G. F. Genetic algorithm for solving the inverse problem of chemical kinetics / G. F. Shaimardanova, K. F. Koledina // Computational mathematics and information technologies. — 2022. — Vol. 1, no. 1. — P. 41–49. Doi 10.23947/2587-8999-2022-1-1-41-49

4. Chistyakov, S. P. Random forests: an overview // Proceedings of the Karelian Scientific Center of the Russian Academy of Sciences. — 2013. — No. 1. — P. 117–136.

5. Brunton S. L., Kutz J. N. Data analysis in science and technology / translated from English by A. A. Slinkin. — Moscow : DMK Press, 2021. — 574 p.

6. Andreas Muller, Sarah Guido Introduction to Machine Learning using Python. O’Reilly. — 2016. — 340 p.

7. Sun S., Huang R. An adaptive k-nearest neighbor algorithm // 2010 seventh international conference on fuzzy systems and knowledge discovery. — IEEE, 2010. — Vol. 1. — P. 91–94.

8. Aurelien Geron Applied machine learning using Scikit-Learn and TensorFlow. — Moscow : Dialectics, 2018. — 690 p.

Received by the editorial office 10.02.2023.
Received after reviewing 02.03.2023.
Accepted for publication 03.03.2023.

About the Authors:

Slepnev, Sergey V., student, Ufa State Petroleum Technological University (1, Kosmonavtov St., Ufa, 450064, RF), ORCID, 14b59b59@gmail.com

Koledina, Kamila F., Dr.Sci. (Phys.-Math.), Associate Professor, Ufa State Petroleum Technological University (1, Kosmonavtov St., Ufa, 450064, RF), researcher, Institute of Petrochemistry and Catalysis, Russian Academy of Sciences (141, Oktyabrya prospect, Ufa, 450075, RF), ORCID, koledinakamila@mail.ru

Claimed contributorship

Slepnev Sergey Vyacheslavovich: program development and algorithm implementation. Koledina Kamila Feliksovna: task formalization, testing, specification development.

Conflict of interest statement

The authors declare that there is no conflict of interest.

All authors have read and approved the final version of the manuscript.