Computers and Artificial Intelligence；Vol.1, No.1; 2024
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/a5ghvb15
Published by: Zeus Press

An Investigation of Molecular Property Prediction and Classification Based on Machine Learning Algorithms

Yanzi Chen

Hangzhou Normal University, Hangzhou 311121, China

Abstract

With the rapid advancement of science and technology, chemical and physical research has entered an era of complexity and high dimensionality, where traditional research paradigms struggle to optimize vast chemical parameter searches. The Machine Chemist platform was developed in this context, leveraging big data and intelligent models to automate chemical synthesis, characterization, and testing processes. This study aims to predict the y_1, y_2 and y_3 properties and classes of 2,580 molecules based on their physicochemical properties and improve model accuracy through data analysis and modeling. Data preprocessing involved removing missing values, duplicates, and outliers using the quartile method, resulting in an analyzable dataset. A scatter plot of y_2 and id suggested a univariate polynomial function relationship, leading to the construction of a univariate nonlinear regression model. The model achieved a high prediction accuracy, with a low root mean square error and a high coefficient of determination. Additionally, the relationship between class and $y_1 \sim y_3$, $x_1 \sim x_{100}$ indicators was examined, revealing mostly nonlinear relationships. A random forest model was established to classify 2,580 molecules into 1 to 4 classes based on the properties of 200,000 chemical molecules. The model's performance was evaluated using decision trees, confusion matrices, precision, and recall metrics. Finally, the SHAP method assessed the impact of feature indicators on classification outcomes, contributing to the development of a reliable molecular category prediction model.

Keywords

model prediction, classification, random forest

1. Introduction

With the rapid advancement of science and technology, research in the fields of chemistry and physics is gradually moving towards a new stage of greater complexity and higher dimensionality. Traditional chemical research methods rely on exhaustive. However, with the continuous expansion of the chemical space and the significant increase in the number of parameters, the traditional methods can only achieve local optimization in the exploration of chemical formulations and process parameters, but not the global optimization of the search.

In this era, the University of Science and Technology of China (USTC) Machine Chemist Platform was born. On this machine chemist platform, the predictive study of the data is crucial. The classes of 200,000 chemical molecules and their 103 physicochemical properties of $y_1 \sim y_3$, $x_1 \sim x_{100}$ and the data of $x_1 \sim x_{100}$ properties of 2580 molecules that need to be predicted are now known. It is now necessary to

construct an accurate mathematical model based on the given data and perform data analysis to help the machine chemist to solve the problem of prediction and classification of 2580 molecules efficiently. The main problems are to investigate whether there is a functional relationship between y_2 and the molecule id, and to try to predict y_2 directly from id, to analyze the data of $y_2 \sim y_3$, $x_1 \sim x_{100}$.

2. Random Forest-based Molecular Classification Model

2.1 Modeling

For analyzing the relationship between class and $y_1 \sim y_3$, $x_1 \sim x_{100}$ metrics, a class prediction model of the molecule is developed to analyze the resulting importance of the feature metrics for classification. Our goal is to determine which features are evaluated to play an important role for categorization in a given set of features $y_1 \sim y_3$ and $x_1 \sim x_{100}$. Considering that the random forest model is capable of constructing multiple decision trees and synthesizing the analysis results of each decision tree, it solves the classification problem of the dataset. In addition, the random forest model is suitable for multivariate nonlinear regression problems, and the multiple variables $y_1 \sim y_3$ and $x_1 \sim x_{100}$ involved in the problem exhibit nonlinear relationships, which is consistent with the properties of the random forest model. Therefore, we choose to build a random forest model to solve this problem in order to comprehensively evaluate the impact of each feature on classification.

2.2 Model solving

2.2.1 Selection of Characterization and Prediction Data

The $y_1 \sim y_3$, $x_1 \sim x_{100}$ properties are used as feature data, and the classification results of CLASS are used as predictive data to find out which feature indicators have a greater impact on the classification results. In order to reduce the number of input feature parameters, a total of 500 feature parameters are selected in this paper. In order to reasonably display the classification results of class, the class category is defined as 1~4. After determining the feature parameters, the training of random forest is carried out through the existing python module library Scikit-learn.

2.2.2 Divide the dataset into training and test sets

In order to test the reliability of the results and the accuracy of the prediction results, the data are randomly divided into training set (70%) and test set (30%)[1]. In order to make the data samples in the test set more representative, Bootstrapping is used to form the training data set T_k. It is known that each feature parameter forms a data set of $y_1 \sim y_3$, $x_1 \sim x_{100}$, and each molecular property we selected corresponds to a T_i, assuming that the data set after sampling is $\{T_1, T_2, \cdots, T_k\}$, each T_i corresponds to one feature set.

2.2.3 Constructing a decision tree based on information entropy and information gain

To evaluate how well the $y_1 \sim y_3$, $x_1 \sim x_{100}$ attributes partition the sample set class 1~4, we use information entropy and information gain metrics to construct decision trees and thus evaluate the category prediction model. Information entropy is a metric used to measure the uncertainty of the sample set class 1~4. In classification problems, the higher the information entropy, the greater the uncertainty of the sample set, that is to say, the more classes are contained in the sample set and the more evenly distributed. Information gain is the amount of reduction in information entropy after the sample set is divided. The greater the information gain, the better the attribute is for the way of dividing the sample set, which can divide the sample set into purer sub-sets.

For the dataset, define the information entropy $H(D)$ as $H(D) = -\sum_{i=1}^{4} p_i \log_2 p_i$, where 4 is the number of categories and p_i is the percentage of category i in the dataset.

For the features, their information gain is calculated by the formula:

$$Gain(T_i) = H(D) - \sum_{j=1}^{m} \frac{|D_j|}{|D|} H(D_j) \tag{1}$$

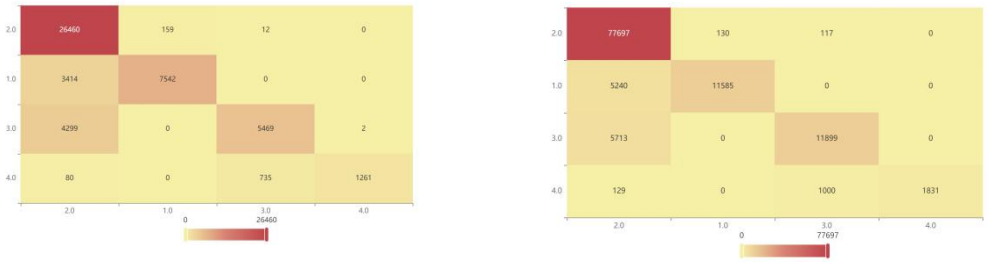
where m is the feature T_i the number of values taken by the D_j is a subdataset divided according to a feature T_i of a certain value taken by a subdataset, and $|D_j|$ is the sub-dataset D_j size of the subdataset. Thus, 4 decision trees are generated.

2.2.4 Training and Evaluation Using Confusion Matrix Models

The five-fold cross-validation method was used to build and train the random forest model to test the accuracy of the prediction results. Through the cross-validation score, the prediction result is, the result accuracy is as high as 99.40%, we determine that the overall prediction ability of the model is more stable and has a certain degree of accuracy. This indicates that the use of the random forest model for class classification is effective and can effectively assess the degree of influence of each variable on class classification.

Next the model is evaluated. Because the confusion matrix is a table used to evaluate the performance of a classification model that compares the predictions of the model with the actual class1 to 4 values. Therefore the heat map of confusion moments for test and training data is plotted as shown below:

Figure 1: Heat map of confusion moments for test set and training set



And the known forecast results are tabulated below:

Table 1: Table of projected results

		Projected results	
		standard practice	negative example
the real situation	real example wigwam	TP (True Positive)	TN (true negative example)
		FP (False Positive)	FN (false negative example)

Where TP denotes the number of positive classes predicted to be positive, TN denotes the number of negative classes predicted to be negative, FP denotes the number of negative classes predicted to be positive and FN denotes the number of positive classes predicted to be negative. Evaluating and training each classifier yields the report shown below :

Table 2: Accuracy, Recall, Precision for Training Set, Cross Validation Set, Test Set and F_1 result table

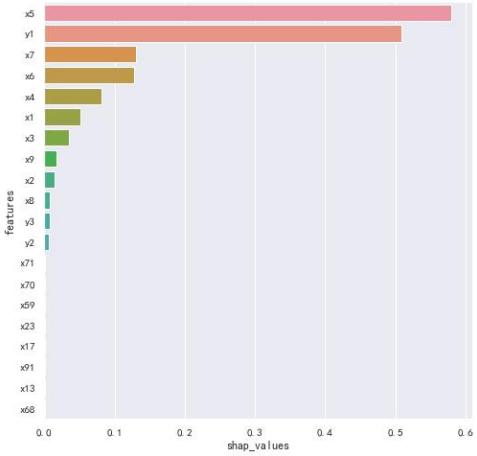
	accuracy	recall rate	accuracy	F_1
training set	0.893	0.893	0.901	0.887
cross validation set	0.872	0.872	0.878	0.86
test set	0.824	0.824	0.849	0.814

Where precision is a measure of the proportion of samples predicted as positive categories that are actually positive categories by the classifier, and recall is a measure of the proportion of all positive category samples that are successfully predicted as positive categories by the classifier. The score is a comprehensive evaluation of the performance of the classifier, which takes into account both Precision and Recall, and is a reconciled average of Precision and Recall. precision, recall, and F1 scores range from 0 to 1, with higher values indicating better performance of the classifiers. As can be seen from the evaluation results, the precision, recall, and F_1 score indices are all close to 1 indicator is very high, indicating that the model predicts well.

2.2.5 Class classification prediction based on shap methods

For classification prediction of CLASS, complex features are extracted for classification prediction by constructing a deep learning model containing SHAP and combining it with a random forest modeler. We obtained the data as shown in Fig by outputting the SHAP value (absolute value) of each feature, the contribution of $y_1 \sim y_3$, $x_1 \sim x_{100}$ to CLASS, and the visualization is shown below:

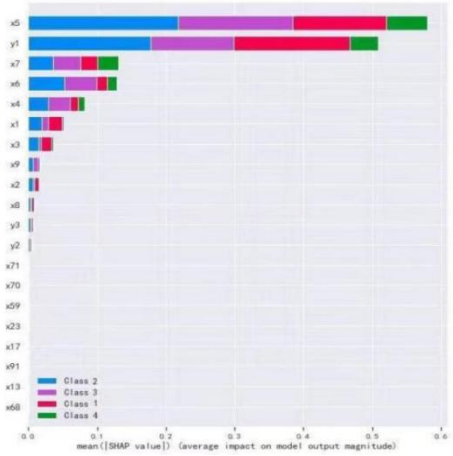
Figure 2: Random forest model based molecular impact on class category prediction table



3. Results and analysis

From the results of the analysis, it can be seen that x_5 has the greatest impact on the category prediction, and the impact is ranked in descending order taking the top 6 as $x_5, y_1, x_7, x_6, x_4, x_1$ respectively. Then according to $y_1 \sim y_3$, $x_1 \sim x_{100}$ different class 1~4 indicators are plotted as shown in the figure below, the results of the analysis can be seen on the degree of influence of class1 in descending order to take the first 6 were $y_1, x_5, x_7, x_1, x_3, x_6$, on the degree of influence of class2 in descending order to take the first 6 were $x_5, y_1, x_6, x_7, x_4, x_1$, on the degree of influence of class3 in descending order to take the first 6 were $x_5, y_1, x_6, x_7, x_4, x_1$, on class4 were $x_5, y_1, x_6, x_7, x_4, x_1$, which indicates that the above indicators have a greater influence on class.

Figure 3 Table showing the impact of Random Forest model-based molecules on the category prediction of the 4 class categories



References

Wang, Lei, Mingyue Chu, Xiaohua Wang, Honglu Guan, Peng Chen & Gao Guanlong. Research on the method of monitoring the operation status of primary side equipment of intelligent substation based on random forest. Electrical Measurement and Instrumentation (07), 184-190. doi:10.19753/j.issn1001-1390.2024.07.026.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

The author would like to thank Ms. Gao for their invaluable guidance and insightful feedback throughout this research.

Copyrights