Machine Learning Methods for Quantitative structure-activity relationship analysis

Helena Dace and Zhenjiao Du

Statistic Department, Kansas State University, Manhattan, KS, 66506, USA.

Contributing authors: helena.newtech@gmail.com; zhenjiao@ksu.edu;

Abstract

Purpose: Machine learning methods would be tested to use specific amino acid descriptor for antioxidant activity prediction, which is expected to the antioxidant peptide QSAR model. **Methods:** We used ten machine learning methods with tuned best hyper-parameters. The cross-validation scores were obtained during hyper-parameters tuning and compared to select the optimal model. **Results:** The Gradient Boosting Regression method got highest cross-validation score on the training data set. Random Forest Regression got the second position. The first seven most important features for the Gradient Boosting Regression model were obtained for demonstration purpose. **Conclusion:** Machine learning methods especially Gradient Boosting Regression and Random Forest Regression can be used to build the QSAR models with higher accuracy than any other previous methods done by researchers.

Keywords: QSAR, Machine Learning, Gradient Boosting Regression, Random Forest Regression, amino acid descriptor, antioxidant activity prediction.

1 Introduction

Amino acids as the building blocks of peptides to a large extent determine the bioactivity of the peptides (Pripp, Isaksson, Stepaniak, Sørhaug, Ardö, 2005). In this research, useful, effective, and low dimension amino acid descriptors will be extracted from the 566 physicochemical properties and used to encode peptides for model development.

Up to now, food protein-derived peptides have gained increasing interest from researchers and consumers (Nongonierma FitzGerald, 2018). Traditional screening on new bioactive peptides is time-consuming and highly relies on many advanced instruments and equipment. QSAR analysis can take advantage of the published data for model development and potential bioactive peptide screening, which will reduce the cost in wet chemistry significantly (Chen et al., 2018; Uno, Kodama, Yukawa, Shidara, Akamatsu, 2020).

Some achievements had been made, such as the prediction of the peptide bitterness and the angiotensin I-converting enzyme inhibitory activity (Nongonierma FitzGerald, 2018; Wu et al., 2014). However, there are a few QSAR studies on the antioxidant activity of peptides and the model performance is also not ideal for practical application.

The best model performance is that R square in cross-validation 0.74 for the ferric thiocyanate (one index of the antioxidant ability) (Chen et al., 2018). For the other indices, the performance was much worse (cross-validation R-square equal to 0.6088 for reducing antioxidant power, 0.630 for oxygen radical absorption capacity, 0.618 for superoxide radical, 0.617 for Trolox-equivalent

antioxidant capacity) (Chen et al., 2018; Deng et al., 2019; Li Li, 2013).

QSAR model is based on the amino acid descriptor to quantitatively describe single amino acids and combines the descriptors of a peptide to predict bioactivity. So far, there are some amino acid descriptors published, such as T-scale and V-scale (Collantes Dunn, 1995; Lin, Long, Bo, Wang, Wu, 2008; Sandberg, Eriksson, Jonsson, Sjöström, Wold, 1998; Tian, Zhou, Li, 2007; Yousefinejad, Hemmateenejad, Mehdipour, 2012). However, no specific amino acid descriptor for antioxidant activity prediction is published, which is expected to the antioxidant peptide QSAR model.

In addition, the machine learning methods such as support vector machine and neural network are shown as useful tools for non-linear analysis, which might have a great potential for the QSAR studies in peptide antioxidant activity prediction (Chen, Chen, Yao, Li, 2018)

This paper proceeds in three sections. In the first section, we describe the ten different machine learning methods that we used to train the model, how we select the optimal model, and data prepossessing and splitting, as well as the analysis methods. Next, we provide information of the results and what significance improvement we have made.The third section is discussion and conclusion about the project.

2 Methods

2.1 Dataset source

A total of 566 numerical indices of amino acids were collected by Beautiful Soup (4.5.3) from AAIndex. Those indices with missing values for amino acids were deleted and the rest indices were 553 in total. Antioxidant tripeptides were manually selected from BIOPEP-UWM and their activities were obtained from the published literatures and expressed as the TEAC values (M TE / M peptides).

2.2 Data processing

2.2.1 Tripeptide encoding and feature selection

The pre-screening numerical indices of amino acids were used to encode tripeptides. Briefly, if there are 'n' numerical indices were selected after the pre-processing, every amino acid could be encoded as a '1 × n' vector. For a tripeptide, it had three amino acids, so it would be encoded as a '1 ×3n' vector where the 1 to the n elements in the vector belonged to N-terminal amino acid, and n+1 to 2n elements referred to the central amino acid and the 2n+1 to 3n elements belonged to the C-terminal amino acid. After the encoding, every tripeptide was represented by 3n variables and the 3n variables would be further screened by Pearson collinearity control method to find independent variables for antioxidant activity prediction as amino acid descriptors for model development.

2.2.2 Collinearity control and standardization

If an absolute value of Pearson's correlation coefficient between two indices was beyond 0.90, one of them would be removed randomly due to the strong correlation. The size of features reduced from 1659 to 386. Then 386 features were standardized for model training.

2.2.3 Data splitting

The total observation for "ABTS" dataset which we were focused on is 130. To avoid bias, we shuffled the rows of the feature matrix (X) and the antioxidant activity (y), at a fixed random state for results reproducing purpose. The training data set is 70 percent of total number of observation (130), and the testing data set is 30 percent.

2.3 QSAR model development

2.3.1 Regression models

Because the target variable antioxidant activity is continuous, we must use regression method rather than classification method. The 10 different models we have tested were support vector machine regression (SVR), random forest regression (RFR), gradient boosting regressor (GBR), tree based XGBoost regressor (tree-XGB), linear based XGBoost regressor (linear-XGB), K-nearest neighbors regression (KNN), partial least squares transformer and regressor (PLSregression), bagging based regression (Bagging), multi-layer perceptron regression (MLP), DecisionTreeRegressor (DTR).

2.3.2 QSAR modeling building and optimization

The model building was done using Python 3.8 in a personal computer (Dell desktop with Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz 2.90 GHz). Models were imported from sklearn and xgboost package. Five folds cross-validation was used in GridSearchCV to tune the hyper-parameters so to find highest accuracy and lower the chance of overfitting for the baseline models. The hyper-parameters with the highest accuracy (cross validation score) on the training data would be selected for each model. The accuracy values from different models would be compared to find the optimal model.

2.3.3 Model performance evaluation

We use the optimal model with the best hyper-parameters to do prediction with held out testing data set and then we compare the R2 scores and mean square errors after prediction on testing data set to those on training data set. Good model for the user would be that both measures for training data set and testing data set are very close to each other. We would also present a list of the features should be selected to get the certain accuracy on the test data. We chose the first seven most important features for the optimal model to demonstrate the further steps.

3 Results

3.1 Results

Among ten QSAR models (Table 1), Gradient Boosting Regression gained the best performance in cross validation 0.84. Random forest regression got the 2nd best performance at cross- validation R2 score 0.80. In the rest models the cross-validation values were below 0.8. Using Gradient Boosting Regression with best hyper-parameters to train the train data set, then predict for the hold out (never used) test data set, we got the train score at 0.98, the test score at 0.74, the train error at 0.043, and the test error at 0.529, which still shows over-fitting. The reason behind this is that the size of data set is too small, with only 130 observations.

3.2 Selected QSAR model performance on the test data set

Table 2 shows the list of the features should be selected to get the certain accuracy on the hold out test data. To achieve test R2 score at 73 percent, seven variables were selected (Table 3) and then were used to encode 130 tripeptides as the matrix of selected features for model (130×7). Based on the variable importance, NADH010102 amino acids contributed the most to the antioxidant activity (y-vector), while FASG760105 amino acids contributed least to the activity.

3.3 Selected features description

Here is the list of descriptions and titles of the first seven most important features selected by the optimal model. (https://www.genome.jp/aaindex/)

NADH010102 D Hydropathy scale based on self-information values in the two-state model (9accessibility) (Naderi-Manesh et al., 2001) T Prediction of protein surface accessibility with information theory

TSAJ990102 D Volumes not including the crystallographic waters using the ProtOr (Tsai et al., 1999) T The packing density in proteins: standard radii and volumes

ARGP820103 D Membrane-buried preference parameters (Argos et al., 1982) T Structural prediction of membrane-bound proteins

CIDH920101 D Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992) T Hydrophobicity and structural classes in proteins

KUMS000102 D Distribution of amino acid residues in the 18 non-redundant families of mesophilic proteins (Kumar et al., 2000) T Factors enhancing protein thermostability

BIOV880102 D Information value for accessibility; average fraction 23T Secondary structure prediction: combination of three different methods

FASG760105 D pK-C (Fasman, 1976)

3.4 Tables

There are four tables shown in the paper. The table of accuracy (cross-validation score) of ten trained models that were trained with the best hyper parameters; the table of selected model's performance; the table of feature numbers recommended for the certain test accuracy; the table

of the first seven most important features for the optimal model.

3.5 Figures

There are five figures shown in the paper. See appendix A. The figures are for the following demonstration: Gradient Boosting performance on training data set; Gradient Boosting performance on test data set; Selected model's deviance update during training and predicting; Permutation importance of the selected model on hold out test data set; The first seven most important features for the selected model.

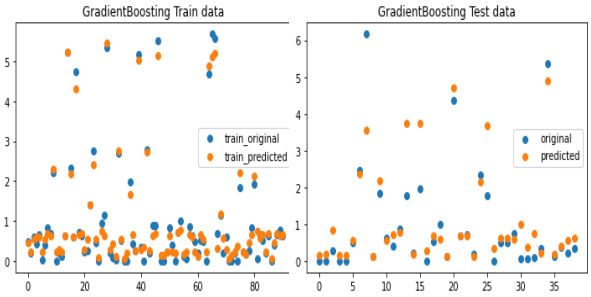

Fig. 1 Training prediction and Testing prediction

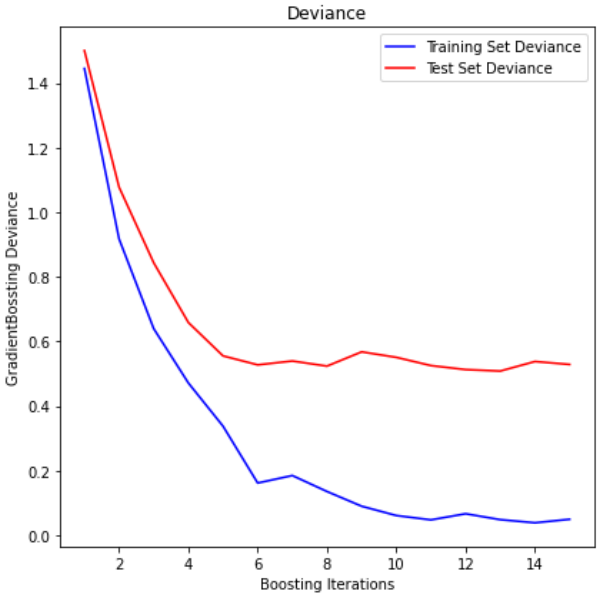

Fig. 2 Selected model's performance on test data set

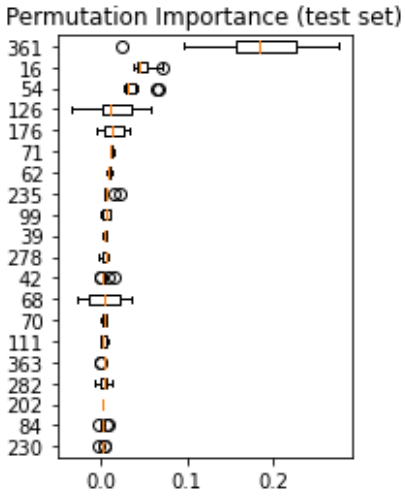

Fig. 3 Permutation importance on test data set

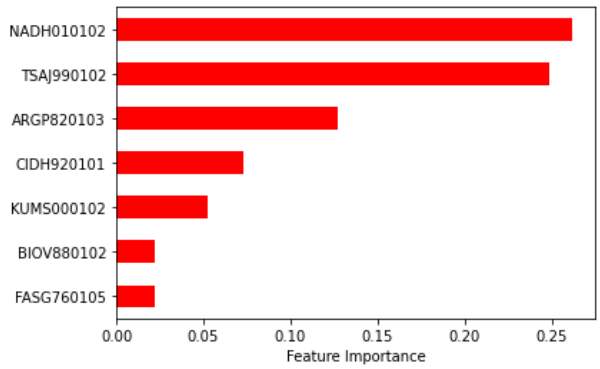

Fig. 4 The first seven most important features

4 Discussion and Conclusion

The final selected features by the user should be practical to the experiments. Conclusions may be used to restate your hypothesis or research question, restate your major findings, explain the relevance and the added value of your work, highlight any limitations of your study, describe future directions for research and recommendations. As we can see, using machine learning method, Gradient boosting regression model gained the best performance in cross validation value 0.84, which has been slightly improved relative to the previous researchers' best score at 0.74. The hyperparameters tuning for different models can be more sophisticated if we have enough time and experience. In the future, we will add more models

Table 1 Selected model's performance

SVR	RFR	GBR	tree-XGB	linear-XGB
0.73	0.8	0.84	0.77	0.28

KNN	PLSR	Bagging	MLPR	DTR
0.67	0.47	0.58	0.74	0.45

Table 2 Selected model's performance

Train score	Test score	Train error	Test error
0.98	0.74	0.04	0.53

in and do wider exploration on hyper-parameters tuning.

Acknowledgments and Authors' contributions

In this paper, Zhenjiao Du has brought the project in and done data collection and data preprocessing. Helena Dace has done the coding and summarizing part of work.

- Availability of data and materials: data for project.zip
- Code availability: code would be provided upon inquires.

References

Chen, N., Chen, J., Yao, B., Li, Z. (2018). Qsar study on antioxidant tripeptides and the antioxidant activity of the designed tripeptides in free radical systems. Molecules, 23(6). https://doi.org/10.3390/molecules23061407

Collantes, E. R., Dunn, W. J. (1995). Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogues. Journal of Medicinal Chemistry, 38(14), 2705–2713. https://doi.org/10.1021/jm00014a022 Deng, B., Long, H., Tang, T., Ni, X., Chen, J., Yang, G., ... Yi, L. (2019). Quantitative structure-activity relationship study of antioxidant tripeptides based on model population analysis. International Journal of Molecular Sciences, 20(4). https://doi.org/10.3390/ijms20040995 Li, Y. W., Li, B. (2013). Characterization of structure-antioxidant activity relationship of peptides in free radical systems using QSAR models: Key sequence positions and their amino acid properties. Journal of Theoretical Biology, 318, 29–43. https://doi.org/10.1016/j.jtbi.2012.10.029

Lin, Z. hua, Long, H. xia, Bo, Z., Wang, Y. qiang, Wu, Y. zhang. (2008). New descriptors of amino acids and their application to peptide QSAR study. Peptides, 29(10), 1798–1805. https://doi.org/10.1016/j.peptides.2008.06.004

Nongonierma, A. B., FitzGerald, R. J. (2018). Enhancing bioactive peptide release and identification using targeted enzymatic hydrolysis of milk proteins. Analytical and Bioanalytical Chemistry, 410(15), 3407–3423. https://doi.org/10.1007/s00216-017-0793-9

Pripp, A. H., Isaksson, T., Stepaniak, L., Sørhaug, T., Ardö, Y. (2005). Quantitative structure activity relationship modelling of peptides and proteins as a tool in food science. Trends in

Table 3 Feature numbers with test accuracy

Thresh=0.022, n=7, Accuracy: 73.62%
Thresh=0.022, n=6, Accuracy: 56.62%
Thresh=0.053, n=5, Accuracy: 58.97%
Thresh=0.073, n=4, Accuracy: 62.54%
Thresh=0.127, n=3, Accuracy: 63.04%
Thresh=0.248, n=2, Accuracy: 63.06%
Thresh=0.261, n=1, Accuracy: 63.27%

Table 4 The first seven most important features

NADH010102	0.261435
TSAJ990102	0.248423
ARGP820103	0.126951
CIDH920101	0.072869
KUMS000102	0.052601
BIOV880102	0.02245
FASG760105	0.022039

Food Science and Technology, 16(11), 484–494. https://doi.org/10.1016/j.tifs.2005.07.003
Saito, K., Jin, D. H., Ogawa, T., Muramoto, K., Hatakeyama, E., Yasuhara, T., Nokihara, K. (2003). Antioxidative properties of tripeptide libraries prepared by the combinatorial chemistry. Journal of Agricultural and Food Chemistry, 51(12), 3668–3674. https://doi.org/10.1021/jf021191n Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. Journal of Medicinal Chemistry, 41(14), 2481–2491. https://doi.org/10.1021/jm9700575 Tian, F., Zhou, P., Li, Z. (2007). T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. Journal of Molecular Structure, 830(1–3), 106–115. https://doi.org/10.1016/j.molstruc.2006.07.004
Uno, S., Kodama, D., Yukawa, H., Shidara, H., Akamatsu, M. (2020). Quantitative analysis of the relationship between structure and antioxidant activity of tripeptides. Journal of Peptide Science, 26(3). https://doi.org/10.1002/psc.3238 Wu, S., Qi, W., Su, R., Li, T., Lu, D., He, Z. (2014). CoMFA and CoMSIA analysis of ACE-inhibitory, antimicrobial and bitter-tasting peptides. European Journal of Medicinal Chemistry, 84, 100–106. https://doi.org/10.1016/j.ejmech.2014.07.015
Yousefinejad, S., Hemmateenejad, B., Mehdipour, A. R. (2012). New autocorrelation QTMS-based descriptors for use in QSAM of peptides. Journal of the Iranian Chemical Society, 9(4), 569–577. https://doi.org/10.1007/s13738-012-0070-y