I COBBIND
Creating connections between biotechnology and industrial sustainability
August 25 to 28, 2024
Costão do Santinho Resort, Florianópolis, SC, Brazil
BIOPROCESS ENGINEERING

GAUSSIAN NOISE INJECTION AS A SUITABLE TECHNIQUE FOR DATA AUGMENTATION IN RANDOM FOREST MODEL APLLIED TO β-GALACTOSIDASE IMMOBILIZATION

Raimundo V. A. Maia[1], Andréa S. Pereira[1] & Luciana R. B. Gonçalves[1*]

[1]Center of Technology, Chemical Engineering Department, Federal University of Ceará, Fortaleza, Brazil.
** email: LRG@ufc.br*

ABSTRACT

In this paper, a dataset was assembled containing 101 observations taken from the literature, correlating the immobilization parameters of the β-galactosidase enzyme with its optimum temperature, optimum pH and the number of cycles in which 60% of the activity is conserved. In order to deal with the small quantity of data to feed and train the model, a copy of the train set was made and random Gaussian noise was injected in it, to simulate variation between experiments, and the two training sets were merged. The performance of the Random Forest model was then evaluated through the use of metrics such as the coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The Random Forest model seem to benefit from this technique, as the R^2 for both the train and test set in all the variables where improved, as well as the MAE and RMSE decreased, a signal of a more well-constructed model. Nevertheless, it is also necessary to explore the possibilities that this technique has to offer in the field of enzymatic immobilization.

Keywords: Immobilization. Gaussian noise. β-galactosidase. Random Forest.

1 INTRODUCTION

β-galactosidase is a well-known glycosidase that is capable of carbohydrates and and pre-biotic galacto-oligosaccharides, present in many substrates, such as milk and dairy products in a cost-effective and environment-friendly way[1]. However, some of the major drawbacks of this enzyme are its thermal and operational stabilities, that can be improved through the immobilization process under optimum pH and temperature conditions. The process of determining such conditions can be costly and experimentally exhaustive, since the plot of a graph is often necessary.

For that reason, machine learning techniques can be used to assist this process, for the model can fed with inputs and outputs from previous experiments and help to determine variables of interest[2] such as the optimum pH, optimum temperature and number of cycles post-immobilization where the enzyme retains 60% of its original activity. One of the challenges of this approach is that the models rely on the quantity and variation between the data points, and it's not possible gather a larger dataset.

One suitable solution for this situation is to artificially simulate variation in the dataset using Gaussian noise injection in a copied training set and posterior merging with the original training data, as the model will have more gain of information and be aware of a wider variety of conditions. Therefore, the objective of this work is to compare the performance of the normal data set and the augmented data set for the abovementioned target variables using a Random Forest model (RFM).

2 MATERIAL & METHODS

The dataset used in this study has 102 observations, divided between 91 points for training and 11 for testing, and was assembled from data extracted from scientific literature from sources such as Science Direct, PubMed and Google Scholar. The input variables to be analyzed by the RFM were: the immobilization method, buffer ionic strength, buffer's ion, enzyme source, concentration of the enzymatic solution, pH of the medium, immobilization temperature and immobilization time. The ambient temperature was considered to be 25 °C and the overnight time was 12 hours, as this is the most reported time for enzymatic immobilization. The output variables were temperature optimum, pH optimum and number of cycles. K-fold Cross Validation[3] and Grid Search[3] were used in order to better train the Random Forest algorithm and search for the better hyper parameters of each one. In addition, all the data was normalized from 0.1 to 0.9, since there were different ranges between the variables due to the units of each.

For the Gaussian noise injection, other works[4] also utilized this technique as a way of augmenting a dataset, although this is a more common practice when applied to image and sound processing. In our case, the randomness of the noise was controlled with a predetermined seed, that generates the same result in all the cases. The equation for the Gaussian probability density function[4] can be seen below:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)}{2\sigma^2}}$$

Where μ represents the mean of *x*, σ the standard deviation and *x* the Gaussian noise to be generated. After the process was made, the training dataset was doubled, and 1456 data points were used to train the RFM.

It is worth noting that algorithms were coded in Python, with the comparison being made according to the values of the coefficient of determination (R^2), mean absolute error (MAE) and root mean squared error (RMSE), where a bigger value is desirable for the first and a lowers values are more representative for the MAE, and RMSE. All the independent variables were used during the training, tuning and test.

3 RESULTS & DISCUSSION

The data before and after the processing of data augmentation can be seen in Figure 1. It can be seen that the distribution of the train set remains practically the same, since the selected values for μ and σ could not alter it so much. The values of R^2 for the train and test set were considerably improved ranging from 0.833 and 0.753 to 0.914 and 0.780, respectively, and MAE, and RMSE for both sets decreased, which indicates that the model is making more accurate average predictions and being less penalized by the presence of larger error coming from outliers.

Table 1 Metrics of performance for temperature optimum (original and augmented train set)

Metric	Train set	Test set	Train set (augmented)	Test set (augmented)
R^2	0.833	0.753	0.914	0.780
MAE (°C)	3.143	2.380	2.335	2.290
RMSE (°C)	4.232	3.586	3.040	3.386

The same seems to apply for the pH optimum. One of the results of artificially generating data points is to prevent the algorithm from overfitting and making it more robust against possible unknown data, since the objective of the Gaussian noise in our case is to emulate a sort of variation between experimental point, resulting in a more reliable model. In this case, the improvement was more subtle, but still can be noticed.

Table 2 Metrics of performance for pH optimum (original and augmented train set)

Metric	Train set	Test set	Train set (augmented)	Test set (augmented)
R^2	0.834	0.745	0.895	0.786
MAE	0.382	0.677	0.357	0.610
RMSE	0.496	0.868	0.455	0.793

As for the number of days, the R^2 for train set was greatly improved, especially for the train set, from 0.765 to 0.851. For this variable, further preprocessing was required to fill the missing values originated from the lack of information available with the median of the distribution (more robust than the mean against outliers). Only the MAE for the augmented train set seems to be higher, but it remains practically the same.

Table 3 Metrics of performance for number of cycles where 60% of enzyme activity is preserved (original and augmented train set)

Metric	Train set	Test set	Train set (augmented)	Test set (augmented)
R^2	0.765	0.726	0.851	0.765
MAE (days)	0.590	0.470	0.605	0.444
RMSE (days)	0.984	0.673	0.853	0.623

4 CONCLUSION

In conclusion, Gaussian noise injection showed to be a suitable data augmentation technique that can create more reliable models when more data is not available, which occurs with great frequency in enzymatic immobilization. For that reason, further exploring is required to refine this technique as to obtain a more realistic simulation of the real world conditions.

REFERENCES

[1] LU, L., GUO, L., WANG, K., LIU, Y., XIAO, M. 2020. Biotech. A. 39. 107465
[2] CHAI, M., MORADI, S., ERFANI, E., ASADNIA, M., CHEN, V., RAZMJOU, A. 2021. Chem. Mat. 33 (22). 8666-8676.
[3] BEHESHTI, N. 2022. Cross Validation and Grid Search. *In:* Towards Data Science.
[4] KANG, Z., FENG, L., WANG, J., 2024. Ind. Chem. Eng. Research. 63. 843-855

ACKNOWLEDGEMENTS

The authors are grateful to UFC (Universidade Federal do Ceará), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil), and FUNCAPES (Fundação Cearense de Apoio ao Desenvolvimento Cientifico e Tecnológico) for the financial support provided.