Symbolic regression for scientific discovery: an application to wind speed forecasting

Ismail Alaoui Abdellaoui
NeuraFirst
Morocco
hello@neurafirst.com

Abstract

Symbolic regression corresponds to an ensemble of techniques that allow to uncover an analytical equation from data. Through a closed form formula, these techniques provide great advantages such as potential scientific discovery of new laws, as well as explainability, feature engineering and fast inference. The present paper aims at applying a recent end-to-end symbolic regression technique, i.e. the equation learner (EQL), to get an analytical equation for wind speed forecasting. We show that it is possible to derive an analytical equation that can achieve reasonable accuracy for short term horizons predictions only using a few number of features.

1 Introduction

A great amount of natural phenomena are explained and computed through a short mathematical expression. For instance, the phenomenon of gravity is explained through Newton's laws of motion. Those equations generally contain a low number of terms and constants, allowing us to understand the relationship between them. Symbolic regression aims at searching the space of mathematical expressions that best fit a given dataset. The main motivation behind this approach is to get an interpretable model that offers an alternative to the black-box models such as neural networks. Thanks to this property, these models have been widely used in industrial empirical modeling Sun et al. (2019); Vázquez et al. (2020). The present paper aims at extending the work in Kim et al. (2020) and applying it to real weather data that includes a high number of input features. Furthermore, we study to which extent can we obtain an analytical equation with a low number of features and terms, while maintaining a reasonable accuracy. More specifically, here we focus on the wind speed forecasting task for three Danish cities. The same dataset as in Mehrkanoon (2019) is used and the obtained results of the proposed model are compared with those of Mehrkanoon (2019).

2 Methodology

The EQL layer architecture is mainly based on a dense layer, yet incorporates some elements designed for symbolic regression Kim et al. (2020). This neural network includes activation functions that act as primitive functions for the final analytical expression. The way these functions operate on their input will depend whether each function is a unary operator (e.g. $\cos(\cdot)$, $\sin(\cdot)$, $(\cdot)^2$, etc.) or a binary operator (e.g. '$+$', '$-$', '$\times$', etc.). The training procedures consists of 2 phases. The first phase uses a particular type of regularization that will be explained in the next section. During the second phase, we first perform a thresholding that will set weights below a certain value at 0. Then the zero-valued weights are frozen while another training occurs that does not make use of regularization. To encourage sparsity and to avoid overfitting, regularization was another key element of the EQL approach Kim et al. (2020). The authors in Kim et al. (2020) used $L_{0.5}^*$, a smoothed version of the $L_{0.5}$, because it empirically led to a more stable gradient descent. More specifically, the weights are

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

computed as follows:

$$L^*_{0.5}(w) = \begin{cases} \|w\|^{\frac{1}{2}} & \text{if } \|w\| \geq a \\ (-\frac{w^4}{8a^3} + \frac{3w^2}{4a} + \frac{3a}{8})^{\frac{1}{2}} & \text{if } \|w\| < a \end{cases}$$

3 Data description

Here we use the same wind speed dataset as used in Mehrkanoon (2019), which is publicly available [1]. The dataset used originates from the National Climatic Data Center (NCDC) and concerns 5 Danish cities and 4 weather features spanning from 2000 to 2010. The time resolution of this dataset is hourly, and the weather features include the temperature, the pressure, the wind speed, and the wind direction.

4 Results and Conclusion

The obtained analytical equation for each target city is tabulated in Table 1. For readability purposes, we replace some repeated parts of the formula that contain the relevant input features by the variables Θ, Δ, and Γ for the cities of Esbjerg, Odense, and Roskilde, respectively. We showed in this paper that the EQL-based approach can yield a mathematical expression that results in an accurate enough wind speed prediction. The major added-value of this approach is its explanability since the compact mathematical expression incorporates the relevant input features as well as their relations. Relevant code used to train and test the algorithms can be found on Github [2].

Table 1: Discovered analytical equation for each target city for 6h ahead wind speed predictions.

City	Analytical expression
Esbjerg	$0.32(1 - 0.08sin(\Theta))^2 - 0.1sin(\Theta) - 0.64sin(2.12sin(\Theta) - 1.37) - 0.83sin(2.61sin(\Theta) + 5.69) - 1.36$
Odense	$-2.99(1 + \frac{0.16}{\Delta})^2 + 0.59 - \frac{1.16}{\Delta}$
Roskilde	$-14.42(-0.20 - \frac{0.10}{\Gamma}) * (0.05 + \frac{0.61}{\Gamma}) - 0.02 * (1 + \frac{0.67}{\Gamma})^2 - 1.21$

References

Kim, S., Lu, P.Y., Mukherjee, S., Gilbert, M., Jing, L., Čeperić, V., Soljačić, M., 2020. Integration of neural network-based symbolic regression in deep learning for scientific discovery. IEEE Transactions on Neural Networks and Learning Systems .

Mehrkanoon, S., 2019. Deep shared representation learning for weather elements forecasting. Knowledge-Based Systems 179, 120–128.

Sun, S., Ouyang, R., Zhang, B., Zhang, T.Y., 2019. Data-driven discovery of formulas by symbolic regression. MRS Bulletin 44, 559–564.

Vázquez, E.V., Ledeneva, Y., García-Hernández, R.A., 2020. Combination of similarity measures based on symbolic regression for confusing drug names identification. Journal of Intelligent & Fuzzy Systems , 1–11.

[1] https://sites.google.com/view/siamak-mehrkanoon/code-data?authuser=0

[2] https://www.github.com/IsmailAlaouiAbdellaoui/EQL-Wind-Speed-Forecasting/