GC2024-STP067

Multistage time series forecasting algorithm based on machine learning methods used to forecast the state of the earth’s magnetosphere

Roman Vladimirov[1] , Vladimir Shirokiy[1] , Oleg Barinov[1] , Sergey Dolenko[1] , Irina Myagkova[1]

[1] D.V. Skobeltsyn Institute of Nuclear Physics, M.V. Lomonosov Moscow State University , Russia

vladimirov@sinp.msu.ru

The idea of this study is the adaptation and application of a 4-step neural network based algorithm for analyzing multidimensional time series to forecast the occurrence of certain events and to identify their precursors - phenomena represented by an unknown combination of parameter values describing the object [1]. Besides forecasting events, the algorithm can be used to forecast the values of some continuous-valued quantities. In this study, the algorithm was applied to forecast the values of the flux of relativistic electrons with E > 2 MeV in geostationary orbit, as well as the values of Dst and Kp geomagnetic indices. The results were compared by relative error and by the sets of the most significant input features selected by the algorithm.

The developed approach allows for adaptively selecting both physical input features and specific delay values when considering the history of each physical feature within the delay embedding approach. This can provide a better understanding of the processes occurring in the studied object (in this case, the Earth's magnetosphere). The following hourly average values were used as input features for the algorithm:

1. Solar wind (SW) parameters at the Lagrange point L1 between the Earth and the Sun: SW velocity ν (km/s), SW temperature T (K), proton density in SW n (cm^{-3}).

2. Interplanetary magnetic field (IMF) vector parameters at the same point L1 in the GSM system: IMF components Bx, By, Bz, IMF magnitude B_{magn} (nT).

3. Geomagnetic indices: equatorial geomagnetic index Dst (nT), planetary geomagnetic index Kp (dimensionless).

4. Random noise features added to test if the system recognizes them as irrelevant for forecasting the target variable.

The set also included four harmonic variables (two with daily period and two with yearly period) to account for rotation of the Earth around its axis and around the Sun (no delay embedding was used for these variables).

The overall scheme of the 4-step algorithm includes the following steps:

1. Selection of the most significant physical features (variables) among those that, in the researcher’s opinion, may affect the predicted target value. An iterative approach is used for this purpose in the current implementation. Within this approach, the system evaluates the correlation (Pearson/Spearman) between the input feature and its delays and the predicted variable. Then, a portion of variables that will be used further is selected based on a specified threshold.

2. Selection of the range of used delays. The considered sets of the input features were created as follows. Set 0 included the selected input variables at the current time. Set 1 included all input variables from set 0 and also all these variables with the delay of 1 time step. Set 2 included all input variables from set 1 and also all these variables with the delay of 2 time steps, and so on up to the researcher-defined limit. A machine learning model is trained on each set created within this cycle, and its quality is evaluated on a pre-held-out dataset. The cycle stops when, based on a specified criterion, increasing the delay range no longer significantly improves the forecasting accuracy.

3. Selection of the most important input features from the obtained 2D feature space limited during the first two stages. Some of the standard approaches to assessing the importance of input features are suitable for this stage.

4. Model hyperparameter tuning.

It is demonstrated that the considered 4-step algorithm can improve the quality of time series forecasting, as well as it can provide an optimal set of input features that allows one to make conclusions on the main interconnections among the target forecast variables and the input features.

Reference

1. S.Dolenko et al. LNCS 5769, 295-304. https://doi.org/10.1007/978-3-642-04277-5_30