Machine Learning Enhanced NARMAX Model Averaging for Forecasting of Geomagnetic Field Indices

Hua-Liang Wei, Simon N. Walker, Richard J. Boynton, and Michael A. Balikhin
Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, UK

1. Abstract

A machine learning enhanced NARMAX model averaging (MLE-NARMAX-MA) approach is proposed for building predictive models for space weather parameters forecasting. Specifically, it aims to build predictive models to characterise the relationship between solar wind parameters and geomagnetic filed indices, for example the dependent relationship of Kp index on many solar wind parameters and magnetic field indices, including solar wind velocity (V), southward interplanetary magnetic field (Bs), solar wind rectified electric field (VBs), and dynamic flow pressure (P).

2. Motivations

In the literature, several classes of statistical machine learning algorithms are now available for sparse model identification. These algorithms include: 1) Orthogonal least squares (OLS) [1]; 2) Orthogonal matching pursuit (OMP) [2]; 3) Least absolute shrinkage and selection operator (Lasso) and its modification version LAR (least angle regression) [3]. The first two classes of algorithms are based on L_2 –norm, whereas the third one is based on L_1+L_2 norm. Each of these algorithms has been successfully applied to solve real-world problems in some specific areas, e.g. Lasso is effective for subset selection for static multivariate regression problems. This study aims to investigate the performance of some of these algorithms for Kp index modelling and prediction. Kp index modelling is a multiple input, nonlinear, dynamic system identification problem involving a great number of correlated, lagged variables and interaction terms. The research interest is not only in predicting the future behaviour but also revealing and understanding the most important predictors and interaction variables. Therefore, the NARMAX model [1], due to its attractive properties, is employed as a modelling platform to conduct all the analysis.

3. System Variables

This study focuses on Kp index (output variable) prediction, which involves a total of 10 solar wind parameters and magnetic field indices (these are input variables). These variable are shown in **Table 1**. The main objective is twofold: 1) build a robust model that can generate 3-hour ahead predictions for Kp index, using hourly measured input variables and 3-hourly measured output variable; 2) find the leading variables and leading interaction variables, i.e., the most important variables and terms for 3-hour ahead prediction for Kp index. Data samples of Kp index are shown in **Fig. 1**.

Table 1 System Input and Output Variables

Name	Description
Kp	Kp index (system output variable; the follows are 10 system input variables)
n	Solar wind density (proton density) [cm^{-3}]
p	Solar wind pressure (flow pressure) [nPa]
V	Solar wind speed (flow speed) [km/s]
Bs	Southward interplanetary magnetic field [nT]
VBs	VBs = V × Bs/1000
$\sqrt{p}$	Square root of p
$\sqrt{V}$	Square root of V
B_x B_y B_z	Three directional components of the interplanetary magnetic field [nT]

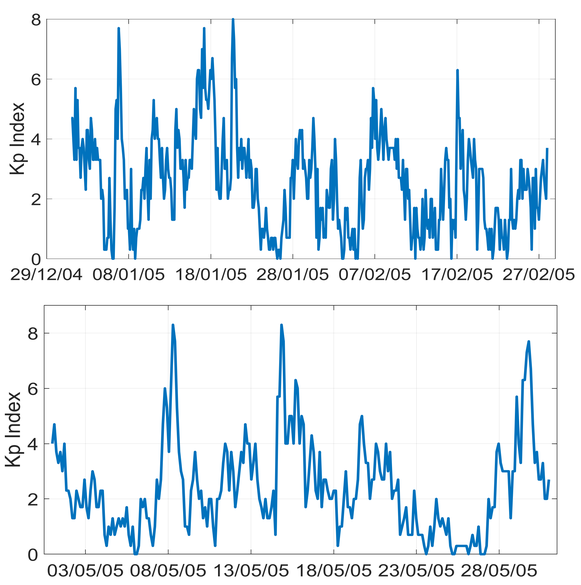

Fig. 1 Sample data for Kp index

4. Methods

The NARMAX (autoregressive moving average with exogenous inputs) model is used to represent the nonlinear dynamic dependent relationship of Kp on the 10 input variables. The model is of the form:

$$y(t) = f(y(t-3), u_1(t-1), u_1(t-2), u_1(t-3), \ldots, u_{10}(t-1), u_{10}(t-2), u_{10}(t-3), e(t-1), \ldots, e(t-3))$$

where $y(t)$ and $u_i(t)$ (i=1,2, …, 10) are output and inputs signals, respectively; and $e(t)$ is the modelling error signal used model parameter filtering at the last modelling stage; $f(\cdot)$ is a nonlinear function which is unknown in advance and can be identified from experimental data by using some model structure detection algorithms. This study employs a distributed lag dictionary learning (DLDL) approach to construct a transparent and parsimonious function $f(\cdot)$ that can well represent the relationship between the input and output data. Let

$$x_1(t) = u_1(t-1), \quad x_2(t) = u_1(t-2), \quad x_3(t) = u_1(t-3), \ldots,$$
$$x_{28}(t) = u_{10}(t-1), \; x_{29}(t) = u_{10}(t-2), \quad x_{30}(t) = u_{10}(t-3), x_{31}(t) = y(t-3) = Kp(t-3),$$

The distributed lag dictionaries used in this study are defined as:

$$D_1(t) = \{1, x_1(t), x_2(t), \ldots, x_{31}(t)\}$$

$$D_2(t) = \left\{ \begin{matrix} x_1(t)x_1(t), & x_1(t)x_2(t), \ldots, & x_1(t)x_{31}(t), \\ & x_2(t)x_2(t), \ldots, & x_2(t)x_{31}(t), \\ & \ldots, \ldots, & x_{31}(t)x_{31}(t) \end{matrix} \right\}$$

Note that the dictionary D=D_1+D_2 contains a total of 528 elements, but not all of them are important for representing and predicting Kp(t). Only a small number of important elements need to be included in the final predictive model.

In this study, the three types of machine learning algorithms, OLS, OMP, Lasso/LARS are used to detect the most important elements from the distributed lag dictionary D, which are used to build sparse models.

5. Results

The data shown in the top panel (1st January–28th February 2005) in Fig. 1 is used as training data to establish models, and the data shown in the bottom panel (1st – 31st May 2005) is used for model performance test. The model performance of four machine learning algorithms for NARMAX model identification based on the specified distributed lag dictionary D is shown in **Table 2**. The performance of the ensemble (model averaging) of the two models obtained by using OLS and iOLS is partly illustrated in **Fig. 2**.

Table 2 A Performance Comparison of Four Algorithms

Algorithm	No of	RNMSE		Corr		PE	
	Terms	Training	Test	Training	Test	Training	Test
LARS	20	0.699	1.262	0.756	0.633	0.512	0.075
OMP	8	0.479	0.529	0.878	0.849	0.771	0.720
OLS	10	0.473	0.524	0.881	0.853	0.777	0.726
iOLS	9	0.499	0.478	0.887	0.879	0.751	0.772

iOLS = iterative OLS; RNMSE= root of normalized mean errors;
Corr = correlation coefficient; PE = prediction efficiency

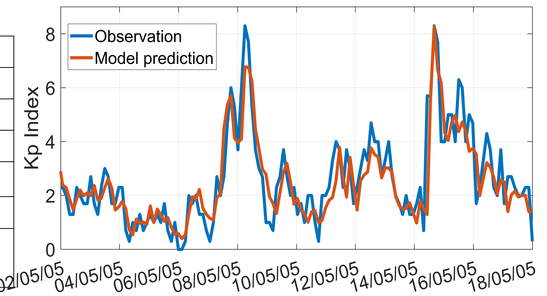

Fig. 2 Model averaging prediction for the period (1-18 May 2005)

6. Concluding Remarks

This study focuses on nonlinear dynamic multiple-input sparse model identification of Kp index using a distributed lag dictionary learning scheme. It needs to deal with a great number of predictive variables and cross-product terms, which are usually highly correlated. The ensemble (model averaging) of the two models produced by OLS and iOLS shows promising robust generalization performance. The performance of LARS, however, appears to be much inefficient for the problem here. This may be because the requirement by LARS and LASSO that the response variable be zero-mean and all regressors be centralized (or normalized in some way), which is inapplicable for complex nonlinear dynamic model identification.

References

[1] S.A. Billings, Nonlinear System Identification, Chichester, UK: Wiley & Son, 2013.
[2] J. Tropp and A. C. Gilbert, “Signal recovery from partial information via orthogonal matching pursuit,” IEEE Trans. Info. Theory, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer: New York, 2001

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 637302.

Contact: Hua-Liang Wei (w.hualiang@sheffield.ac.uk)