Random Forest, Support Vector Regression and Gradient Boosting Methods for Ionosphere Total Electron Content Nowcasting problem at mid-latitudes

Aleksei Zhukov[1], **Denis Sidorov**[1] [2], **Anna Mylnikova**[1], **Yury Yasyukevich**[1] [3]
(1) Institute of Solar and Terrestrial Physics of Russian Academy of Sciences, Irkutsk, Russia zhukovalex13@gmail.com
(2) Institute of Energy Systems of Russian Academy of Sciences, Irkutsk, Russia
(3) Irkutsk State University, Irkutsk, Russia

Abstract

This paper illustrates data-driven machine learning approach for ionosphere total electron content (TEC) forecasting. The authors exploit different state-of-the-are machine learning algorithms like random forest, support vector regression, and gradient boosting to achive high accuracy (higher than conventional naive and linear models). The proposed approach allows to determine the most important parameters. The approach revealed that current TEC, first time derivative of TEC, cosine from local time LT, current F10.7 and SYM/H indexes, exponential moving averages of TEC (with 12, 24, 96 hour periods), 12h-lagged, 2-days and 15-days lagged F10.7 are the significant features for vertical TEC 4-hour nowcasting model. As the experimental data, the vertical absolute TEC was used. The time resolution of the data is 30 minutes. Initial phase and psueduorange slant TEC were recorded by the mid-latitude station IRKJ (52 N, 104 E) in 2014. All the models were evaluated and testing results comparison provided. Machine learning based models allow us to achive small RMSE $\approx$ 3 TECU, linear regression model based on significant features results in $\approx$ 4.5 TECU, while naive models results to huge RMSE.

1 Introduction

The ionospheric parameters nowcasting is a quite urgent task for radio communication and radar systems, control and positioning systems [1], [2]. The dynamics of ionospheric parameters have become increasingly important to forecast. One of the parameters is total electron content (TEC) which can be used for ionosphere correction in radio systems [3].

TEC forecasting and nowcasting leads us to the complex mathematical problem of prediction in multi-dimentional feature space. There are a lot of machine learning approaches to solve such problems. The main of them it is neural networks [5] and machin learnning approaches: Random Forest (RF) [9], Support Vector Regression (SVM)[7] and Trees Gradient Boosting (TGB) [8] Methods . In this paper we focuses on two important aspects of this problem: selection of nowcasting input features and building of the final model. We used Random Forest, Support Vector Regression and Gradient Boosting Methods. Proposed models exploit machine learning to approximate TEC variations. We consider 4-hour interval for nowcasting.

2 Input data and feature selection

As input data we used absolute vertical TEC data deduced from dual-frequency combined phase and pseudorange GNSS measurements [4]. Along with I_V we obtained first and second time derivatives of TEC. To obtain TEC and its derivatives wee used GPS/GLONASS station IRKJ (52 N, 104 E) data for 2014. In order to provide additional information about solar and geomagnetic activity F10.7, SYM/H and AE indices were employed.

Also we used an apriori knowledge about periodical nature of considered time series, since we added cosine from local time ($cos(2\pi \cdot LT/24)$) to feature list, where LT is a local time in hours. By analogy with classical time series analysis methods moving average and autocorrelation analysis were exploited. Thus using correlation analysis for feature extraction we assume that the highest parts of cross-correlation function (specific lags) should be more informative. In this way we used data lagged by 0.5, 12, 24, 48, 125, 360 hours. We used also exponential moving averaging (2, 3, 4, 12, 24, 48, 72, 96 hours) from initial data as input parameters.

3 Modelling results

RF model was exploited to evaluate a relevance of input parameters. Results on relevance are presented in Fig. 1. With adherence to this values recursive feature eliminations was used to get set of the most relative parameters. As one can see the main parameters for 4-hours ahead regression model are current TEC, first time derivative of TEC, cosine from local time LT, current F10.7 and SYM/H indexes, exponential moving averages of TEC (with 12, 24, 96 hour periods), 12h-lagged, 2-days and 15-days lagged F10.7.

Fig. 2 shows result on 4-hour nowcasted TEC along with target data (black curve). Random Forest, Trees Gradient Bosting and SVM results are shown in red, green and blue,

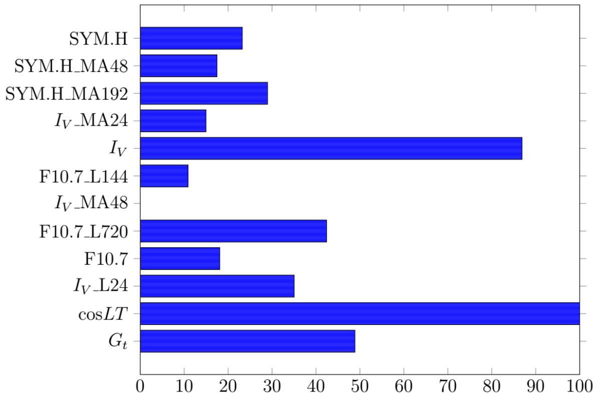

Figure 1. Random forest evaluated Input parameters relevance. Where I_v - TEC values, G_t - TEC time derivate, MA means moving average with specific period, L denotes lagged features, cosLT is a cosine from local time.

correspondingly. The main problem as one can see is TEC nowcasting during night-time. The night-time relative error can exceed 100%.

Root-mean-square error (RMSE) and mean average error are shown in Tab. 1. All the values are in TECU units. RF and TGB RMSEs are $\tilde{3}$.5 and 3.3 TECU, while $\tilde{4}$.5 for SVM with linear kernel. MAE is $\tilde{2}$.5 for RF.

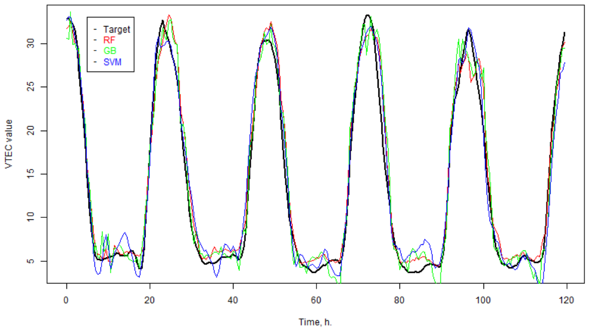

Figure 2. RF, TGB and SVM models predictions (2014-11-24 till 2014-11-29).

We also constructed three naive models: first use a current value as a forecast (N1), second exploits first time derivative to get evaluation (N2). As a third model multiparameter linear regression built with selected parameters (LR). Tab. 1 shows that both N1 and N2 models for 4-hour nowcasting don't allow appropriate result obtaining, but LR RMSE are almost the same as SVM those.

Table 1. Predictive models error comparison fro 4-hour forecast.

	RF	TGB	SVM	N1	N2	LR
RMSE	3.49	3.30	4.49	9.14	16.51	4.61
MAE	2.49	2.35	3.50	6.74	12.23	3.66

In order to compare the influence of forecasting period on model error all evaluated models were tested with different periods. Results can be seen in fig.3. It is clear that for all the models the error (RMSE) increases with increase in forecasting period. There is sharp increase in RMSE at 0-2 hour interval. After 2 hour RMSE increases more slowly. SVM results is almost the same as those for multiparameter linear regression.

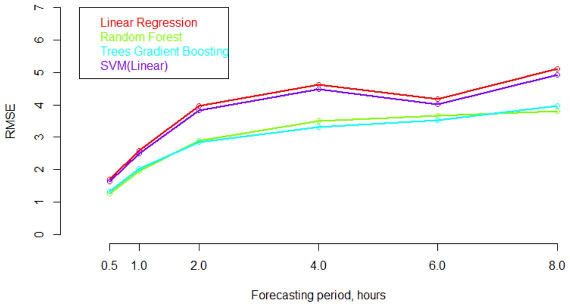

Figure 3. Error of forecasting models for different periods.

4 Conclusions

Problem of ionosphere total electron content nowcasting is considered in terms of the state of the art machine learning methodology. The special attention is paid to the problem of feature selection.

Machine learning technique allows to select relevant parameters for effective TEC nowcasting. The approach revealed that current TEC, first time derivative of TEC, cosine from local time, current F10.7 and SYM/H indexes, exponential moving averages of TEC (with 12, 24, 96 hour periods), 12h-lagged, 2-days and 15-days lagged F10.7 are the significant features for vertical TEC nowcasting model. Based on Random Forest and recursive feature eliminations we find the most important parameters for TEC nowcasting.

As it was shown even simple linear model fitted with relevant parameters gave quite a good results. Random Forest and Trees Gradient Boosting machine learning models based on that parameters has rather small RMSE on real data ($\approx 3.4 - 4.5$ TECU) than conventional linear model. So that the multiregression linear regression model (LR) based on selected parameters reproduce duirnal TEC variations in a good way. However its RMSE is $\approx$ 4.5 TECU. As one can see linear SVM model shows low accuracy. The accuracy is comparable with multiparameter linear regression. This leads us to the conclusion that a different SVM kernel should be used. Unfortunately proper kernel can be found only empirically.

Constructed models can be used for TEC nowcasting in radar and radio communication systems for ionosphere effect reduction. Current parameters can be used for creating new advanced models at midlatitudes.

5 Acknowledgement

The authors are grateful to A.M. Vesnin and S.V. Syrovatsky for assistance in the preliminarydata collection. We acknowledge the Space Physics Data Facility (SPDF) data and International GNSS Service IGS [6] for F10.7, SYM/H, AE and GNSS data. This work was performed under the Russian Science Foundation Grant No. 17-77-20005. Initial absolute vertical TEC data were obtained under program of the fundamental research of the state academies of Russia for 2013-2020 years (project No. II.16.1.1).

References

[1] N. Jakowski, A. Wehrenpfennig, S. Heise, I. Kutiev, "Space weather effects on transionospheric radio wave propagation on 6 April 2000," *Acta Geodaetica et Geophysica Hungarica*, vol. 37, no. 2-3, pp. 213 – 220, 2002.

[2] B. Zolesi, L.R. Cander, "Ionospheric Prediction and Forecasting,", Springer Berlin Heidelberg: Berlin, Heidelberg. 2014.

[3] E. Afraimovich and Y. Yasukevich, "Using GPS-GLONASS-GALILEO data and IRI modeling for ionospheric calibration of radio telescopes and radio interferometers," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 70, no. 15, pp. 1949 – 1962, 2008.

[4] Y. Yasyukevich, A. Mylnikova, and A. Polyakova, "Estimating the total electron content absolute value from the GPS/GLONASS data," *Results in Physics*, vol. 5, no. Supplement C, pp. 32 – 33, 2015.

[5] J.B. Habarulema, L.-A. McKinnell, B.D.L. "Opperman,Regional GPS TEC modeling; Attempted spatial and temporal extrapolation of TEC using neural networks, " *Journal of Geophysical Research: Space Physics*, vol. 116, no. A4, A04314, doi: 10.1029/2010JA016269, 2011.

[6] J.M. Dow, R.E. Neilan, C. Rizos, "The International GNSS Service in a changing landscape of Global Navigation Satellite Systems", journal, " *Journal of Geodesy*, vol. 83, no. 3, pp. 191 – 198, doi: 10.1007/s00190-008-0300-3, 2009.

[7] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997.

[8] R. Schapire, "The boosting approach to machine learning: an overview." *Nonlinear estimation and classification, Springer, New York*, pp. 149–171, 2003.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.