

Comparison of outliers and novelty detection to identify ionospheric TEC irregularities during geomagnetic storm and substorm

Asis Pattisahusiwa, The Houw Liong, and Acep Purqon

Physics of Earth and Complex Systems, Institut Teknologi Bandung, Indonesia

E-mail: asis@pattisahusiwa.com

Abstract. In this study, we compare two learning mechanisms: outliers and novelty detection in order to detect ionospheric TEC disturbance by November 2004 geomagnetic storm and January 2005 substorm. The mechanisms are applied by using ν -SVR learning algorithm which is a regression version of SVM. Our results show that both mechanisms are quiet accurate in learning TEC data. However, novelty detection is more accurate than outliers detection in extracting anomalies related to geomagnetic events. The detected anomalies by outliers detection are mostly related to trend of data, while novelty detection are associated to geomagnetic events. Novelty detection also shows evidence of LSTID during geomagnetic events.

1. Introduction

It is well known that geomagnetic storm and substorm, as the effect of interaction between solar and geomagnetic field, contribute to some irregularities in the ionosphere [1, 2]. Several publications show a strong relationship between geomagnetic storm and variabilities in Total Electron Content (TEC) data [2–5]. However, recent review related to the characteristics of disturbances in TEC data indicates that more robust method is still needed to understand the event [6].

On the other hand, Support Vector Machine (SVM) is well known as a quiet accurate method in machine learning. As a learning method, SVM can be used in classification, clustering, time-series analysis, or regression task. Several applications of SVM in ionospheric physics research can be found, for example, in [7, 8].

In this brief study, we compare two learning mechanism, novelty and outliers detection based on ν -Support Vector Regression (ν -SVR) method, which is a regression version of SVM, in order to detect the irregularities in TEC data. Typically, both mechanisms are used to provide input data to unsupervised learning algorithm in machine learning. However, the mechanisms can also be applied in regression task. The result of this study can be taken into consideration for further uses of the method in space weather research as well as on machine learning itself.

2. Data and method

Several data were used to analyze the effect of geomagnetic storm and substorm in ionospheric TEC. Disturbance Storm Time (Dst) and Auroral Electrojet (AE) indices are used as the



indicator of geomagnetic activities. Dst and AE indices are provided by World Data Center (WDC) for Geomagnetism, Kyoto University (<http://wdc.kugi.kyoto-u.ac.jp>). According to Dst and AE indices, we analyzed irregularities in TEC data during geomagnetic storm, 8–10 November 2004, and substorm, 21–22 January 2005. Hence, TEC data were taken from 7–11 November 2004 and 20–23 January 2005 at Global Ionospheric Map (GIM) coordinate 2.5° latitude and 95° longitude. TEC data were extracted from GIM provided by Center for Orbit Determination in Europe (CODE) analysis center and can be accessed at <http://igsceb.jpl.nasa.gov>. All data is analyzed with sampling every 2 hours.

Irregularities in TEC data during geomagnetic storm and substorm were analyzed by using ν -SVR. ν -SVR is a regression version of SVM which is first proposed by Schölkopf in 1999 and developed from Vapnik's ϵ -SVR [9]. Difference between both regression methods is only on determination of ϵ parameter. In ν -SVR, ϵ parameter will be tuned automatically based on ν value. The accuracy of ν -SVR is dependent on ν , regularity constant C , and kernel function. In this study, We use $\nu = 0.5$, $C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|)$, and Radial Basis Function (RBF) as the kernel function,

$$f(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right); \quad \gamma = \frac{1}{2\sigma}; \quad \sigma = 0.2 \times \text{range}(\vec{x}),$$

where \bar{y} , σ_y are mean and standard deviation of label vector; x is feature vector.

Anomalies in TEC data were extracted by using two learning mechanisms, Novelty and Outliers detections. Novelty detection requires noise-free label vector in training stage, while the other one can contain noises. In novelty detection, monthly median TEC and TEC data were used as ν -SVR label vectors for training and prediction stages, respectively; while in outliers detection, TEC data were used for both stages. Dst and AE indices were used as feature vectors for both methods.

TEC irregularities were determined by applied a boundary condition to the prediction result. The data are categorized as the outliers if it meets the following criteria,

$$\begin{aligned} \delta y_i &< \mu_{\delta y} - k\sigma_{\delta y} \\ \delta y_i &> \mu_{\delta y} + k\sigma_{\delta y} \end{aligned} \quad (1)$$

where $\delta y_i = f_i(\vec{x}) - y_i$, $f_i(\vec{x})$ is the prediction result, while $\mu_{\delta y}$ and $\sigma_{\delta y}$ are mean and standard deviation of the error (δy) respectively, and k is a constant.

3. Result and discussion

According to figure 1, ν -SVR method based on both mechanisms show quiet accurate in learning TEC data. Accuracies of outliers detection are shown in figures 1a and 1b, while novelty detection are shown in figures 1c and 1d. Based on the figures, both mechanisms show less accurate on 10 November 2004 where TEC data are more disturbed than the other days. Mean Square Error (MSE) and correlation coefficient values for novelty detection on this day are 230.92 and 0.79 respectively, while for outliers detection are 7.195 and 0.92, respectively. On outliers detection, average MSE for geomagnetic storm and substorm are 8.0102 and 2.1527, while average correlation coefficient are 0.9587 and 0.9893, respectively. Furthermore, the average MSE for novelty detection are 84.6024 and 19.0930, while average correlation coefficient are 0.8940 and 0.9817, for both case respectively.

On figure 2, we can see the difference of prediction result (solid line) for each mechanisms on both case. Prediction result by outliers detection are shown in figure 2a for geomagnetic storm and 2b for geomagnetic substorm. Moreover, prediction result by novelty detection for geomagnetic storm is shown in figure 2c, while geomagnetic substorm is shown in 2d. On the figures, it seems that the outliers detection mechanism is inherently just follows the trend of the

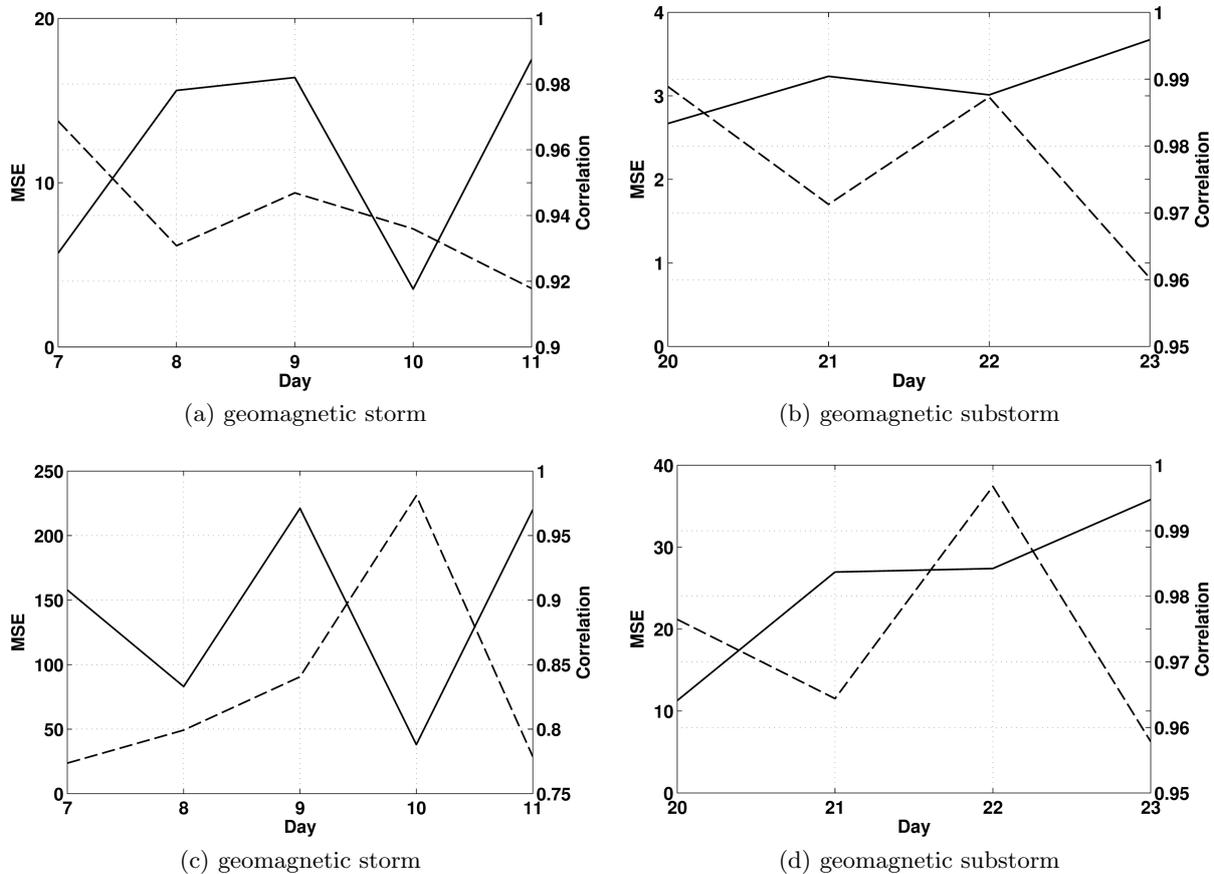


Figure 1: Accuracy of outliers (1a, 1b) and novelty detection (1c, 1d): MSE (dashed line), and correlation coefficient (solid line).

data. Whereas, prediction by novelty detection seems just follows monthly median TEC data. Hence, the result itself gives an advantage to determine anomalies in TEC data.

Obviously, figures 2b and 2c show that the first magnetic Storm Sudden Commencement (SSC) occurred at 7 November 2004 14:00 UTC and followed by main phase geomagnetic storm at 7 November 2004 20:00 UTC. The second main phase began at 9 November 2004 10:00 UTC when the first storm in recovery phase. The figures also show the disturbance of ionospheric TEC on 9–10 November 2004. Similarly, figures 2b and 2d show increasing geomagnetic activities at 21 January 2005 18:00 UTC.

However, only novelty detection is quiet accurate to detect irregularities in TEC data on 9 and 10 November 2004. The irregularities that are detected on 9 November 2004 associated with positive phase of ionospheric storm, while 10 November 2004 associated with negative phase of ionospheric storm. These anomalies also show the evidence of Large Scale Traveling Ionospheric Disturbance (LSTID) caused by geomagnetic storm.

On the contrary, the detected anomalies by outliers detection are look like not related to disturbances by geomagnetic event. Especially before and when first storm onset on 7 November 2004. However, the disturbances by storm onset will be delayed to arrive at equatorial region [1]. Hence, can be concluded that the detected anomalies mostly related to trend of the data.

Similarly, the detected anomalies on figure 2b by outliers detection is associated with trend of data. Whereas, the detected outliers on 22 January 2005 by novelty detection are associated

with increasing geomagnetic activities on 21 January 2005. However, the detected anomalies on 20 and 23 November 2005 by novelty detection are false positive.

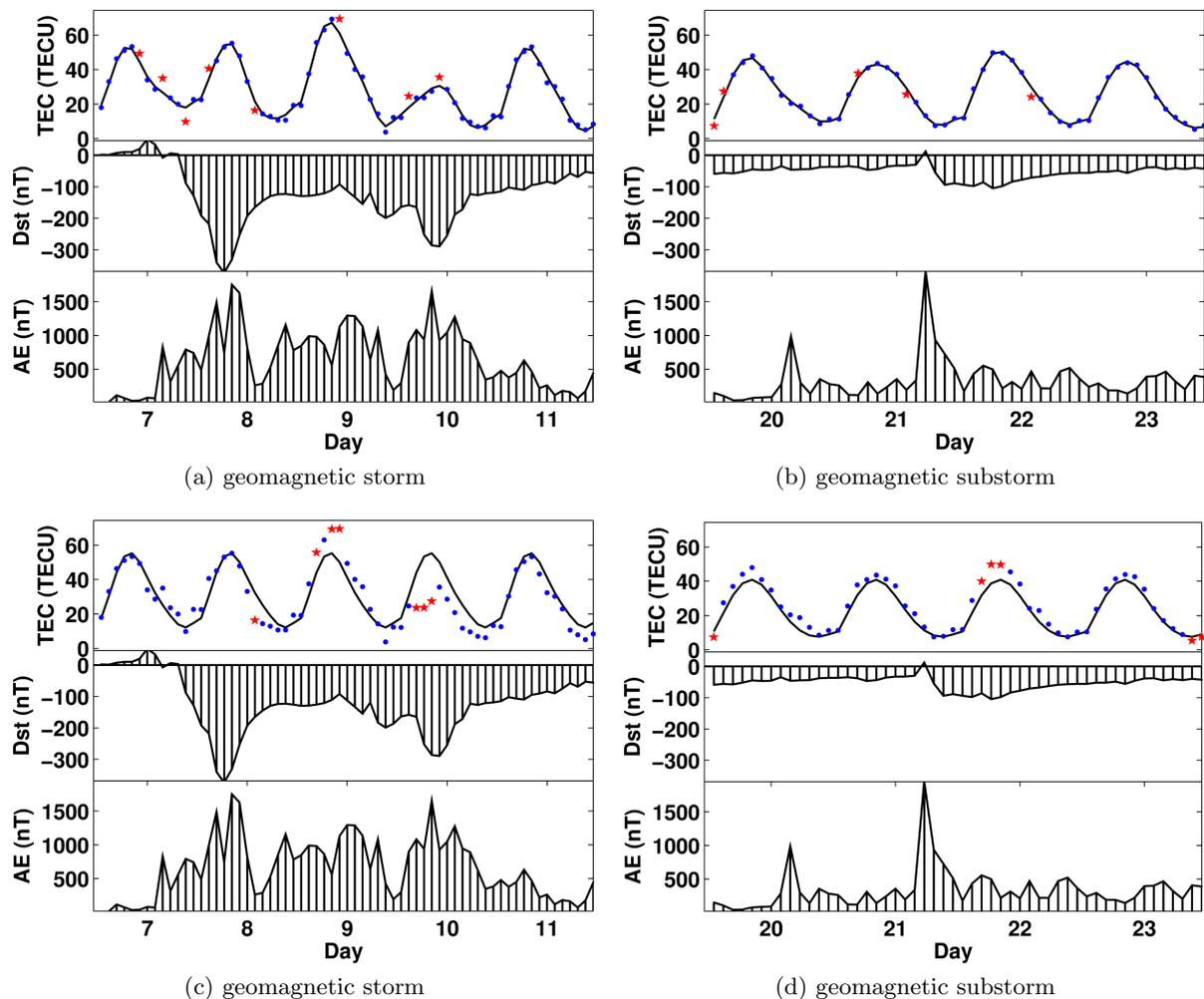


Figure 2: Prediction result of outliers (2a, 2b) and novelty detection (2c, 2d). On TEC graphic: prediction (solid line), TEC (blue circle), and outliers (red star).

4. Conclusion

Based on the result, we can conclude that both mechanisms are quite accurate in the case of learning process. However, novelty detection is more accurate to detect the irregularities related to geomagnetic storm and substorm than outliers detection. The irregularities detected by novelty detection is related to geomagnetic events, while outliers detection is mostly related to trend of the data.

References

- [1] Balan N, Alleyne H, Walker S, Reme H, McCrea I and Aylward A 2008 *Journal of Atmospheric and Solar-Terrestrial Physics* **70** 2101–2111
- [2] Gómez L, Sabbione J I, Van Zele M A, Meza A and Brunini C 2007 *Journal of atmospheric and solar-terrestrial physics* **69** 955–968

- [3] Kim H, Clauer C, Deshpande K, Lessard M, Weatherwax A, Bust G, Crowley G and Humphreys T 2014 *Journal of Atmospheric and Solar-Terrestrial Physics* **114** 1–8
- [4] Arslan N and Demirel H 2007 The effects of geomagnetic storms on ionosphere and gps signals (Turkish national geodetic commission scientific meeting, METU)
- [5] Yamamoto A, Ohta Y, Okuzawa T, Taguchi S, Tomizawa I and Shibata T 2000 *Earth, planets and space* **52** 1073–1076
- [6] Mendillo M 2006 *Reviews of Geophysics* **44**
- [7] Akhoondzadeh M 2013 Support vector machines for tec seismo-ionospheric anomalies detection *Annales Geophysicae* vol 31 (Copernicus GmbH) pp 173–186
- [8] Pattisahusiwa A, Liong T H and Purqon A 2015 *AIP Conference Proceedings* **1677** 060009
- [9] Schölkopf B, Bartlett P L, Smola A J and Williamson R 1999 *Advances in neural information processing systems* 330–336