Symposium on Big Data in Finance, Retail and Commerce: Statistical and Computational Challenges
Lisbon, 2-3 November 2017, Portugal

Forecasting on Imbalanced Big Time Series Using Ensembles of Classifiers and Minimum Information: Application to Ozone Data

Álvaro Gómez-Losada

European Commission, Joint Research Centre (JRC), Economics of Climate Change, Energy and Transport, Edificio Expo, C/ Inca Garcilaso 3, 41092 Seville, Spain, alvaro.gomez-losada@ec.europa.eu

Abstract

Ozone (O_3) is a transboundary air pollutant of major concern on both the European and global scale. At ground level, ozone can be formed by photochemical reactions between anthropogenic nitrogen oxides and volatile organic compounds in the presence of sunlight. O_3 concentrations reveal a strong spatial and temporal variability, with a significant differentiation between rural, suburban and urban sites. The target value for O_3 established by the European Union Air Quality Directive [1] for the protection of human health and vegetation is frequently exceeded in large regions of Europe, including urban agglomerations.

Air quality forecasters often use all available meteorological variables to forecast O_3. Variables retrieved from air quality agencies or meteorological services typically include temperature, solar radiation, surface wind speed and direction, cloudiness and vertical temperature gradient until certain height in atmosphere, to cite a few. However, on many occasions such information is not readily available. Equipment at meteorological or air quality monitoring stations can be limited, or its maintenance programs not duly observed. As a consequence, the data provided are scarce, of poor quality, or both. Furthermore, when variables' availability for forecasting O_3 is not ensured, many of the forecasting models found in literature are not reproducible.

The field of air quality is just one of many that utilizes forecasting models within a data classification framework. Traditional classification algorithms perform poorly on imbalanced data sets [2]. These algorithms fail to properly

represent the distributive characteristics of the data and as a result, provide unfavourable accuracies across the different classes. Also, usual metrics used to evaluate the performance of classifiers in imbalanced learning are not appropriate since the misclassification error of the minority class is far costlier than that of the majority class.

More advanced forecasting models use ensembles of classifier approaches. Generally speaking, an ensemble of classifiers combines the prediction of individual classifier to often produce more accurate results than that of any single classifier. This is due to the ability ensembles have to produce generalization, which is usually higher than that of base classifiers. The computational cost of ensembles is higher than individual learners, but also the model complexity.

The aim of this study is evaluate the O_3 forecasting ability of nine base classifiers and several derived ensembles, using a reduced number of predictors and long time series with severe class imbalance. These predictors were extracted using the time of collecting O_3 hourly concentrations at monitoring stations as the only information; no other additional information related to the O_3 genesis was used.

O_3 hourly concentrations from five monitoring stations from three urban environments in Andalusia (Spain) were obtained for this study, from 2005 to 2015, allowing a range of local atmospheric conditions to be considered. This study adopts the O_3 forecast problem using a binary classification task, where the positive (minority) class included all the instances which values exceeded the limit value set for O_3 according to the Directive 2008 [1], and the negative class, otherwise. In the studied time series from the monitoring sites, the proportion of instances of the minority class to the total of instances ranged from 0.8% to 3.2%.

The model selection was performed after tuning the parameters of the base classifiers using a 10-fold cross-validation scheme. The strategy for dealing with the unbalanced classification tasks was to randomly under-sample the majority class in the training sets before learning the base classifiers. Afterwards, the decision boundary for the positive and negative classes was estimated using the test set and the final forecasting results obtained using the validation set. To evaluate the classification accuracy of ensembles and base classifiers, three specific metrics for measuring performance in class imbalanced learning were used, namely, area under the curve (AUC), precision-recall area (PRA) [3] and Matthew's correlation coefficient (MCC). Additionally, to study the learning curve of base classifiers and ensembles, base classifiers were trained on subsets of different size obtained from initial training

data sets, preserving in them the class imbalance from the original training data. The classification accuracy was later evaluated on the original validation sets using the three metrics indicated above.

The classification accuracy for base classifiers ranged from 0.78 to 0.95 (AUC), 0.78 to 0.94 (PRA) and 0.14 to 0.42 (MCC). The ensemble algorithms achieved similar results, from 0.68 to 0.95 (AUC), 0.63 to 0.96 (PRA) and 0.08 to 0.51 (MCC). Methodology designed shows that no base classifier uniformly outperforms over all the time series studied at different locations, and that gaining accuracy of ensembles is not significant. However, the learning curves of base classifiers and ensembles indicated a faster convergence of the base classifiers. In addition, they indicated that similar accuracy is obtained when training base classifiers, and therefore ensembles, using training data set sizes smaller than those used initially.

Keywords: Classification, time series, imbalanced learning, ensemble methods, forecasting.

Disclaimer

The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

References

[1] Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe [Internet]. [cited 2017 Jul 13].

[2] He, H., Ma, Y. (2013) *Imbalanced learning. Foundations, algorithms and applications*. IEEE Press-Wiley, Canada.

[3] Saito, T., Rehmsmeier, M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced data sets. *Plos One* 10(3), 1-21.