ECML PKDD Discovery Challenges 2017

Roberto Corizzo[1] and Dino Ienco[2]

[1] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
roberto.corizzo@uniba.it
[2] Irstea, UMR TETIS, Univ. Montpellier, France
LIRMM, Montpellier, France
dino.ienco@irstea.fr

1 Introduction

The ECML PKDD Discovery Challenge 2017 has hosted two different tasks: the *Multi-Plant Photovoltaic Energy Forecasting Challenge* challenge and the *Time Series Land Cover Classification Challenge*. The two challenges have involved many participants coming from different parts of the world and they have also contributed to supply real world data to the ECML PKDD community, on which develop and apply their methods. More in detail, the *Multi-Plant Photovoltaic Energy Forecasting* challenge has involved 34 participants from more than 15 countries and more than 15 institutions. Among all participants, 11 submitted their solution to the final round. Considering the *Time Series Land Cover Classification* challenge, it has involved 58 participants from more than 15 countries. Among all participants, 21 submitted their solution to the final round. Due to the big interest arisen by the two challenges, we have decided to collect the description of the best approaches to produce this proceeding. More in detail, we gave the possibility to the top-50% ranked teams, for each challenge, to describe their methods through a research paper. Limiting the submission to the top ranked teams gave us a criteria to gather together the best quality material.

In the rest of the Discovery Challenge Preface we introduce, for each challenge, the general context/outline, the data description and the strategy we set up in order to evaluate the different team results.

2 Multi-Plant Photovoltaic Energy Forecasting Challenge

2.1 Challenge outline

The increasing presence of renewable energy sources has deeply affected the energy market, which is characterized by distributed power generation.

Integrating the contribution of intermittent energy sources, such as photovoltaic (PV) plants, into the smart grid, is a non trivial task. In the literature, this motivated the development of predictive algorithms, with the aim to forecast the energy produced by renewable energy plants for a given time horizon. The adoption of accurate algorithms is indeed fundamental to face the main challenges introduced by the new energy market, such as grid integration, load balancing and energy trading.

In this Discovery Challenge, the task proposed was that of power forecasting for multiple photovoltaic (PV) plants distributed in the south of Italy and connected to a power grid. More specifically, given the historical time series data including the weather conditions and the energy produced for a set of plants, and given only the weather conditions for a future period, the task was to predict the time series of the energy produced for each plant and for the whole period, at hourly granularity.

2.2 Data description

The provided dataset[1] consisted of time series data regarding weather conditions and production collected by sensors on three closely located PV plants in Italy. More in detail, the dataset contained Plant features (Irradiance, Temperature, Power) and Weather features (Cloud cover, Pressure, Wind speed, Dewpoint, Temperature, Humidity and Windbearing).

Plant data have been gathered locally from sensors available on plants, whereas weather data have been queried from Forecast.io (`http://forecast.io/`). Moreover, the irradiance for the test data, for which we are interested in predicting the power, have been queried from PVGIS (`http://re.jrc.ec.europa.eu/pvgis/apps4/pvest.php`). Forecast.io data comes from a wide range of data sources, which are statistically aggregated to provide the most accurate forecast possible for a given location. PVGIS data makes use of monthly averages of daily sums of global and diffuse irradiation. The averages represent the period 1981-1990. The PVGIS database has been queried for each day and plant separately using a custom-made wrapper, specifying the date and the coordinates of the plants of interest (latitude and longitude).

Each time point is the hourly aggregation obtained as the average of all the measures available in a specified hour.

For each plant, day, and variable (such as temperature, irradiance, cloud cover, etc.), data consisted of a time series of 19 values representing hourly aggregated observations (plants are active from 02:00 am to 08:00 pm).

The plants considered in the dataset had a maximal power rate of 1000 kW/h.

It is worth to mention that the dataset available in the challenge was a subset of a bigger dataset, consisting of 17 plants and data observed in the time period from January 1st, 2012 to May 4th, 2014. This dataset has been sent to all the participants at the end of the challenge.

2.3 Issues

The data was affected by the following issues:

- i) **Missing data**, which are related to sensor failures or communication problems (e.g.: zero production observed even if the actual irradiance is positive, or the opposite). There are different ways to address this issue, such as using external data sources given the spatial coordinates of the plants (which have not been revealed for privacy reasons) and the time point, or replacing the null values with hourly averages observed in the past. The participants had the freedom to choose the strategy they preferred;

- ii) **Outliers**, which are related to measurements errors (e.g.: plants considered in the dataset had a maximal power rate of 1000 kW/h, but some dirty measurements were above 400,000). One possible solution consists in identifying abnormal values on the whole dataset and replace them with averages. However, it could happen that some values which are generally considered normal, constitute an outlier for the specific time point characterizing the hour of the day at which they are observed. This allows for more sophisticated data repair stategies;
- iii) **Temporal autocorrelation**, induced by the cyclicity of weather phenomena. In fact, analyzing variables related to weather phenomena it can be observed that: i) they tend to have similar values at a given time in close days, ii) they have a cyclic and seasonal (over days and years) behavior, iii) they tend to show the same trend over time (e.g. summer days are featured by an increased irradiance compared to winter days). When temporal autocorrelation is taken into account, for example resorting to time window models or directional statistics, it can contribute to obtain more accurate prediction models;
- iv) **Spatial autocorrelation**, induced by the spatial proximity of sensors. This leads to the violation of the usual assumption that observations are independently and identically distributed (i.i.d.). This phenomenon is usually taken into account resorting to spatial statistics which consider the spatial dependencies in the learning process. Although the explicit consideration of these dependencies brings additional complexity, it generally leads to increased accuracy of learned models.

From this characterization it looks clear that data preprocessing tasks such as data cleaning and repair should be performed in order to remove dirty measurements of sensors and missing data which could degrade model performances. Moreover, feature extraction mechanisms could help to catch the spatial and temporal autocorrelation in data and improve the predictions.

2.4 Evaluation

Training data consisted of a temporal period of 12 months (year 2012) including the daily target time series (power observed for each plant), whereas testing data consisted of 3 months (January to March 2013) for which the target time series (power) has not been provided.

Considering the issues discussed above, we have suggested to first perform a data cleaning phase in order to remove outliers which could negatively affect the predictive model.

Another issue that had to be considered was the normalization of the power variables, which represents a convention in the power forecasting field that also makes the results in terms of error more understandable afterwards. Therefore we suggested to the participants to perform a min-max scaling on training and testing data, before applying the prediction model, by identifying the maximum value of the power and using the same range $[0, max]$ to normalize all power variables.

The performance score for each participant has been automatically calculated by the system, according to the standard Root Mean Square Error (RMSE) measure.

Five temporary submissions and one final submission have been granted to each participant, in the time frame from Jul 20 to Jul 24 at 23:59:59 (UTC-12 Time Zone). The temporary leaderboard has been automatically obtained on a fixed validation set.

For both the temporary and definitive leaderboards, the ranking has been obtained in ascending order of RMSE, considering only the best submission for each participant.

The reference website for challenge, which includes the dataset, the the instructions and the leaderboard is available at the following URL: `http://193.204.187.201:8080/pv_challenge_website/`.

3 TiSeLaC : Time Series Land Cover Classification Challenge

3.1 Challenge outline

Modern Earth Observation programs produce huge volumes of remotely sensed data every day. Such information can be organized in time series of satellite images that can be useful to monitor geographical zones through time. How to efficiently manage and analyze remote sensing time series is still an open challenge in the remote sensing field.

In the context of land cover classification, exploiting time series of satellite images, instead of one single image, can be fruitful to distinguish among classes based on the fact they have different temporal profiles.

The objective of this challenge is to bring closer the Machine Learning and Remote Sensing communities to work on such kind of data. The Machine Learning community has the opportunity to validate and test their approaches on real world data in an application context that is getting more and more attention due to the increasing availability of SITS data while, this challenge offers to the Remote Sensing experts a way to discover and evaluate new data mining and machine learning methods to deal with SITS data. The challenge involves a multi-class single label classification problem where the examples to classify are pixels described by the time series of satellite images and the prediction is related to the land cover of associated to each pixel. A more detailed description follows.

3.2 Data Description

The dataset has been generated from an annual time series of 23 Landsat 8 images acquired in 2014 above the Reunion Island (2866 X 2633 pixels at 30 m spatial resolution), provided at level 2A. Source data have been further processed to fill cloudy observations via pixel-wise multi-temporal linear interpolation on each multi-spectral band (OLI) independently, and to compute complementary radiometric indices (NDVI, NDWI and brightness index - BI). A total of 10 features (7 surface reflectances plus 3 indices) are considered for each pixel at each timestamp.

Reference land cover data has been built using two publicly available dataset, namely the 2012 Corine Land Cover (CLC) map and the 2014 farmers' graphical land parcel registration (Régistre Parcellaire Graphique - RPG). The most significant classes for the study area have been retained, and a spatial processing (aided by photo-interpretation) has also been performed to ensure consistency with image geometry. Finally, a pixel-based random sampling of this dataset has been applied to provide an almost balanced ground truth. The final reference training dataset consists of a total of 81 714 pixels distributed over 9 classes.

The source data are provided by the French Pôle Thématique Surfaces Continentales - THEIA and preprocessed by the Multi-sensor Atmospheric Correction and Cloud Screening (MACCS) level 2A processor developed at the French National Space Agency (CNES) to provide accurate atmospheric, environmental and geometric corrections as well as precise cloud masks. Data pre-processing and temporal gap filling have been performed using the iota2 [3] Land Cover processor developed by CESBIO[4].

3.3 Evaluation

Training data consisted of 81 714 pixels with their corresponding multidimensional time series (23 timestamps with 10 dimensions each) with associated classification, whereas testing data consisted of 17 973 pixels with their corresponding time series. In order to evaluate the results submitted to the challenge we used the F-Measure averaged over all the classes. More in detail, the F-Measure has been aggregated by a weighted average considering as weight the number of examples for each class. The reference website for the challenge, which includes the dataset, the instructions and the leaderboard is available at the following URL: `https://sites.google.com/site/dinoienco/tiselc`.

References

1. Michelangelo Ceci, Roberto Corizzo, Fabio Fumarola, Donato Malerba, and Aleksandra Rashkovska. Predictive modeling of PV energy production: How to set up the learning task for a better prediction? *IEEE Trans. Industrial Informatics*, 13(3):956–966, 2017.

[3] `http://tully.ups-tlse.fr/jordi/iota2.git`

[4] `http://www.cesbio.ups-tlse.fr/`