WUR – SPECIAL REPORT NO 19 – 2019, 51-56

One-class classification for blight risk forecasting

PETER SKELSEY

The James Hutton Institute, Dundee DD2 5DA, Scotland

SUMMARY

Information from crop disease surveillance programs provide real-world data about the drivers of infection events. This study explored the application of machine learning techniques to survey data for potato late blight outbreaks to derive models for forecasting the risk of disease. Five different anomaly-detection algorithms were compared according to their accuracy in forecasting outbreaks: *k*-means, Gaussian mixture model, kernel density estimator, one-class support vector machine, and isolation forest. The results showed that one-class support vector machine had the highest accuracy followed by Gaussian mixture model, with 98% and 97% respectively. There was added value in combining the algorithms in an ensemble to provide a more robust forecasting tool. The techniques used here can easily be applied to outbreak data from other crop pathosystems to derive tools for agricultural decision support.

KEYWORDS

Phytophthora infestans, late blight, machine learning, anomaly detection, decision support

INTRODUCTION

Models for forecasting risk of crop disease are typically derived from experiments in the laboratory, glasshouse or controlled environment chambers, or derived from statistical relationships between environmental data and disease observations in small plot field trials. An argument can be made, however, that experiments conducted under highly standardized conditions or at a small spatial or temporal scale can produce results with little validity beyond the specific environment in which the experiment is conducted. Plant disease epidemics occur at the landscape-scale and are influenced by aspects of the wider environment, such as topography, climate, and variability in host and pathogen populations. Landscape-scale experiments to derive and test models of disease risk are rarely conducted, but surveillance programs for crop diseases can provide useful real-world data about the drivers of infection events at the spatial scale of interest.

Many decision support systems for crop protection are used to make binary decisions about whether to apply a crop protectant (De Wolf and Isard 2007). This approach can be framed as a classification problem, where input data are used to discriminate between risk and no-risk of disease. In a conventional classification problem, data from each class of interest are available, i.e., healthy and diseased crops, and each object is represented by a vector of features (e.g., date, location, weather data, crop characteristics). Training the classifier involves feeding training examples from both classes to the algorithm so that it can create rules to assign membership based on the individual feature values. When encountering new instances, they can be classified according to the learned rules. In one-class classification, the problem is to classify data when information is available for only one class of observations (Désir *et al.* 2013; Khan and Madden 2010; Mazhelis 2006; Tax 2001). This is an interesting problem because this is typically the case for surveillance programs for crop diseases, where only outbreaks are recorded and data from surrounding healthy crops is omitted. The task, then, is to find the feature values most commonly associated with disease occurrence and those that are anomalous, in order to derive rules that can be used to predict risk/no risk of disease in the future. One-class classification is therefore often considered as outlier or anomaly detection.

In this study we applied anomaly detection algorithms to national late blight survey data to learn the 'normal' weather conditions most commonly associated with outbreaks of disease, and evaluated their ability as tools to forecast risk.

MATERIALS AND METHODS

Input data

The late blight outbreak data spanned a 6-year period (2012–2017) and consisted of the date and coordinates (UK postcode district centroid) of 1049 late blight outbreaks from across GB. These data are collected routinely each year by blight scouts as part of the Agriculture and Horticulture Development Board (AHDB) Potatoes 'Fight Against Blight' campaign. Hourly weather data corresponding to every outbreak location were provided by the UK Met. Office. A 28-day period prior to the date that each outbreak was reported was considered sufficient for relating weather conditions for infection to the dates at which disease was first observed in the crop. On each day in that 28-day period, the minimum daily temperature and the total number of hours of relative humidity $\geq$ 90% were calculated and used as input features (variables) for the anomaly detection algorithms. These two features were selected as they are historically used to calculate late blight risk alerts in GB. The Smith Period was developed in the 1950's as a DSS to indicate high risk periods for PLB development in GB (Smith 1956). This was replaced by the Hutton Criteria in 2017, which is the current national warning system for late blight in GB via the online service 'Blightwatch'. The Smith Period and the Hutton Criteria will serve as baseline models for comparison with the anomaly detection algorithms.

Machine learning

Among the many anomaly detection algorithms available, five of the most commonly used were selected to compare for forecasting outbreaks: one-class *k*-means, Gaussian mixture model, kernel density estimator, one-class support vector machine, and isolation forest (Goldstein and Uchida 2016). Each algorithm assigns an anomaly score to each data point, describing its outlier tendency. A user-defined decision threshold is applied to the returned scores to create a decision boundary separating nominal samples from anomalies. In the current context, any weather conditions that fall inside the boundary trigger a disease risk alert, otherwise no alert is issued. Each algorithm contains parameters that are learned from the data, and 'hyperparameters' whose value must be set before the learning process begins. In order to fairly evaluate the performance of the algorithms and not introduce biases in the choices of hyperparameters, a nested cross-validation scheme was implemented. The procedure consisted of two nested *k*-fold cross-validation loops: an outer one to test generalization accuracy, and an inner one for optimization of hyperparameters and the decision threshold used to separate nominal data (producing a risk alert) from anomalies. Model performance was assessed in both the inner and outer loops using an extrinsic performance function, where the models were used to forecast outbreaks in the held-out test set. Any risk alert triggered in the 28-day window of weather data preceding each outbreak was classed as a 'successful' forecast. The decision threshold value was tuned so that approximately 6 in 7 data points were classified as anomalies, producing a risk alert on every 1 in 7 days, on average. Cross-validation accuracy was then measured as the percentage of outbreaks successfully forecasted. The two models used as a baseline for comparison (the Smith Period and the Hutton Criteria) did not require training or tuning of hyperparameters and were tested for accuracy using the same held-out test sets as the anomaly detection algorithms.

RESULTS

All five optimized anomaly detection algorithms produced similar decision boundaries in the temperature and relative humidity data (Figure 1). There were, however, slight differences in the range of values classified as 'normal'. The daily minimum temperature values defining the decision boundaries ranged from: 9.6-12.2, 9.4-12.1, 9.1-12.2, 9.4-12.1, and 8.0-13.0°C for one-class *k*-means, Gaussian mixture model, kernel density estimator, one-class support vector machine, and isolation forest, respectively. The number of hours of RH≥90% defining the decision boundaries ranged from: 6.0-12.3, 4.6-9.6, 5.4-11.0, 3.5-10.0, and 5.6-11.9 for one-class *k*-means, Gaussian mixture model, kernel density estimator, one-class support vector machine, and isolation forest, respectively.

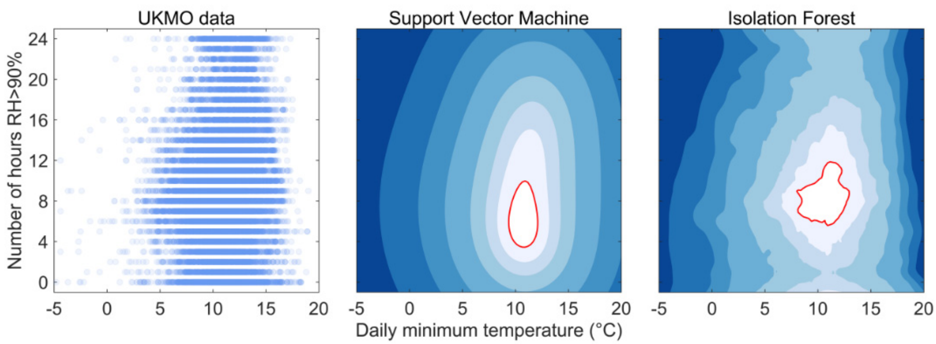

Figure 1. *Comparison of decision boundaries for two of the five optimized anomaly detection algorithms applied to the complete potato late blight outbreak data set: (a) raw weather data in the 28 days preceding each outbreak, (b) one-class support vector machine, and (c) isolation forest. Darker contours correspond to greater outlier tendency, with a bolder line for the decision boundaries*

The Smith Period achieved an accuracy of 81.8% in forecasting outbreaks, with a frequency of alerts of 12.6% (i.e., approximately 1 alert in every 8 days). The Hutton Criteria had a much higher accuracy, at 96.1%, but at the expense of a high frequency of alerts of 31.4% (i.e., 1 in every 3 days). For the anomaly detection algorithms, predictive accuracy was: 95.7, 97.1, 93.8, 98.1 and 96.2% for one-class *k*-means, Gaussian mixture model, kernel density estimator, one-class support vector machine, and isolation forest, respectively. Frequency of alerts was consistent at approximately one alert in every 7 days: 14.9, 12.8, 15.0, 14.7, and 13.7% for one-class *k*-means, Gaussian mixture model, kernel density estimator, one-class support vector machine, and isolation forest, respectively. There was in general good agreement among the anomaly detection algorithms on the dates on which alerts were issued (Figure 2). The Smith Period missed several outbreaks, whereas the Hutton Criteria tended to issue alerts over multiple consecutive days. In order to increase confidence in the robustness of predictions, the five anomaly detection techniques were also combined in an 'ensemble', where each model had a 'vote' on whether an alert should be issued on any given day, and the majority vote won.

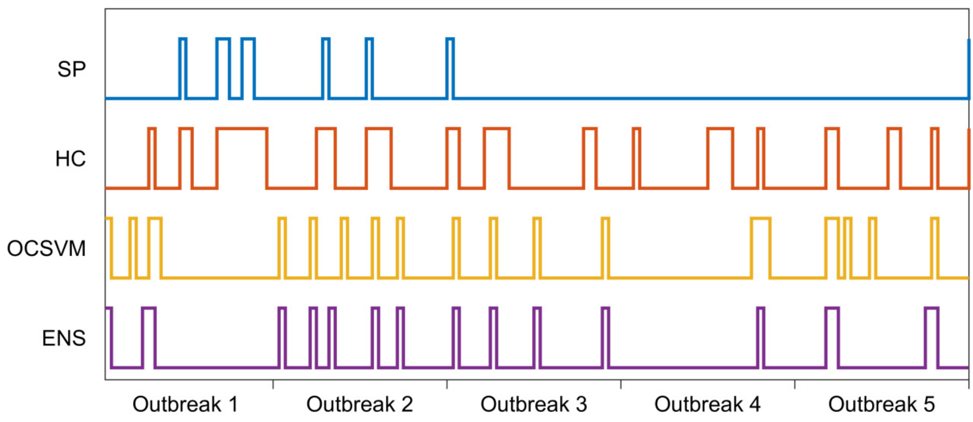

Figure 2. *Comparison of the dates on which alerts were issued in the 28 days leading up to a reported outbreak for a random sample of five outbreaks. Results are provided for the two baseline models (SP = Smith Period, HC = Hutton Criteria), an example anomaly detection algorithm (OCSVM = one-class support vector machine), and an ensemble of the five anomaly detection techniques (ENS)*

CONCLUSIONS

This study outlines a method for deriving crop disease forecasting tools from survey data that contain information on diseased crops only. Anomaly detection algorithms were used to learn the envelope of weather conditions most commonly associated with outbreaks, and then tested for their ability to forecast outbreaks using held-out test data. The algorithms were able to correctly forecast late blight outbreaks with an accuracy of 96.3%, on average, whilst maintaining a low frequency of alerts. The five algorithms used encompassed a range of density, distance, classification, and rule-based techniques to delineate normal/anomalous regions in the input data. Combining these different approaches in an ensemble of anomaly detectors yields a robust tool for forecasting late blight outbreaks that mitigates any weaknesses inherent in the individual models. In future work the ensemble will be made available to the GB potato industry as a cutting-edge 'intelligent decision support system'. The chief advantage of this system over 'conventional' decision support tools will be its ability to continue learning and improving as new outbreak data become available.

ACKNOWLEDGEMENTS

The author is grateful for financial support for this work from the Rural and Environment Science and Analytical Services (RESAS) Division of the Scottish Government under its Environmental Change and Food, Land and People Strategic Research Programmes (2016–2021). The author also thanks AHDB Potatoes and the 'Fight Against Blight' scouts for providing the late blight outbreak data.

REFERENCES

De Wolf, E., and Isard, S.A. (2007). Disease cycle approach to plant disease prediction. Annual Review of Phytopathology 45:203-220.

Désir, C., Bernard, S., Petitjean, C., and Heutte, L. (2013). One class random forests. Pattern Recognition 46:3490-3506.

Goldstein, M., and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE 11:e0152173.

Khan, S.S., and Madden, M.G. (2010). A survey of recent trends in one class classification. Pages 188-197 in: Artificial Intelligence and Cognitive Science, L. Coyle and J. Freyne, eds. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mazhelis, O. (2006). One-class classifiers: a review and analysis of suitability in the context of mobile-masquerader detection. South African Computer Journal 2006:29-48.

Smith, L.P. (1956). Potato late blight forecasting by 90 per cent humidity criteria. Plant Pathology 5:83-87.

Tax, D.M.J. (2001). One-class classification. Delft University of Technology.