Scientific workflow management with ADAMS: building and data mining a database of crop protection and related data

P.M. Reutemann and G. Holmes

Computer Science Department, University of Waikato, Hamilton, New Zealand
Corresponding author: fracpete@waikato.ac.nz

Abstract

Data mining is said to be a field that encourages data to speak for itself rather than “forcing” data to conform to a pre-specified model, but we have to acknowledge that what is spoken by the data may well be gibberish. To obtain meaning from data it is important to use techniques systematically, to follow sound experimental procedure and to examine results expertly. This paper presents a framework for scientific discovery from data with two examples from the biological sciences. The first case is a re-investigation of previously published work on aphid trap data to predict aphid phenology and the second is a commercial application for identifying and counting insects captured on sticky plates in greenhouses. Using support vector machines rather than neural networks or linear regression gives better results in case of the aphid trap data. For both cases, we use the open source machine learning workbench WEKA for predictive modelling and the open source ADAMS workflow system for automating data collection, preparation, feature generation, application of predictive models and output generation.

Keywords machine learning, data mining, weka, workflow, adams.

Introduction

Data mining has many techniques for extracting knowledge from data, such as classification, regression, clustering and association rule mining. Its methods can be applied in many different disciplines. Within data mining methodology, researchers have a vast array of approaches at their disposal that range from simple data mining methods, such as decision trees, to more sophisticated ones, such as support vector machines. WEKA (Hall et al. 2009) is an open source, cross-platform machine learning workbench written in Java and was developed at the University of Waikato. The workbench offers many of these techniques through an easily accessible graphical user interface. Both data exploration and statistical experimentation are possible using this software. In this study we demonstrate the use of the ADAMS scientific workflow system (Reutemann & Vanschoren 2012) that supports rigorous experimentation, captures process for repeatability and deployment of resulting models. A workflow allows a user to define and annotate each step of data collection, preparation, processing, evaluation and output or graph generation, through a graphical user interface. System capability is demonstrated using two case studies. The first case study uses a scientific workflow from a research perspective while the second shows how to move from research to commercial application using the system.

Materials and methods

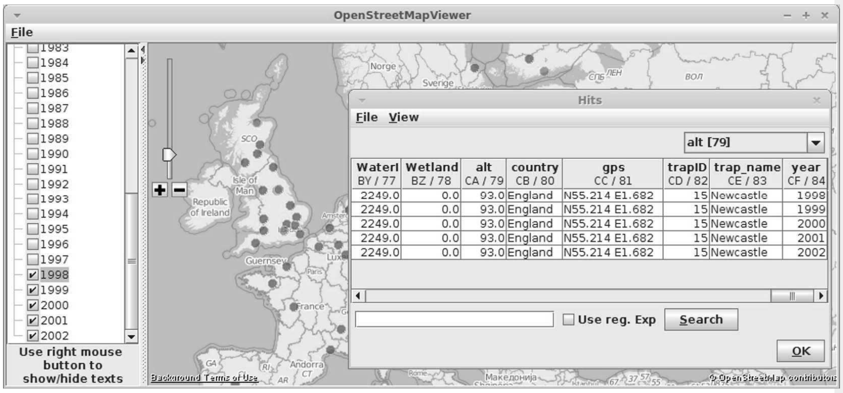
Various data mining techniques are applied to a crop protection database to extract information to stimulate future research and to demonstrate potentially informative patterns that can be extracted from the data.

The crop protection data used in this study comprises trap catches for *Myzus persicae* in a large number of locations throughout Europe and has been previously discussed in detail in Cocu et al. (2005). The data comprises aphid counts gathered by a European network of suction traps, obtained from the EXAMINE database (http://www.rothamsted.ac.uk/insect-survey), and enriched with geographical and climatic variables. Table 1 gives an overview of the variables and Figure 1 shows a portion of the data visualized with an ADAMS workflow, using its geographic information system (GIS) capability.

Table 1. Overview of variables. All areas are in ha with a circle of R = 75km.

Variable	Description
trap_name, trapID	Trap identification
lat, long, alt	Latitude, longitude and altitude associated with the trap
year	Year the data were collected (1969-2002)
XRn, ***X***PRn	Mean rainfall in month ***X***; collection year, previous year
YTmp, ***Y***PTmp	Mean temperature in month ***Y***; collection year, previous year
C***n***	Mean temperature of the coldest ***n*** days
C***n***P***m***	Mean temperature of the next ***m*** days after the coldest ***n*** days
ConFor, DecFor, MixFor, Grass	Coniferous, deciduous, mixed forest, grass land
WaterI, Urban, Sea	Inland waters, urban areas, sea
ArableR, CropPer	Arable land, permanent crops
Shrub, Barren, Frozen, Wetland	Shrubland, barren land, frozen areas, wetlands
MpeJd5thFlt	Julian date of 5th aphid caught in trap (Mpe = *Myzus persicae*)
MpeLgTotW52	$\log_{10}$ (n+1) transformed annual numbers of aphids

For our experiments, we develop models to predict the Julian date that the 5th aphid was caught for timing prediction (“MpeJd5thFlt”) and the $\log_{10}$-transformed annual numbers of aphids (“MpeLgTotW52”) for aphid abundance. All records that contain missing values were removed, as well as trap name and ID variables. Similar to Cocu et al. (2005), we use multiple linear regression (LR) and neural networks (multi-layer perceptron using backpropagation (MLP) with various parameter settings. However, we further evaluate M5' model trees (Quinlan 1992, Wang & Witten 1997), Gaussian processes (GPD) (Rasmussen & Williams 2006) and support vector machines for regression (SMOreg) (Shevade et al. 1999) to model the data. To run the experiments, we use the WEKA Experimenter as supplied by the ADAMS system.



□ 1983
□ 1984
□ 1985
□ 1986
□ 1987
□ 1988
□ 1989
□ 1990
□ 1991
□ 1992
□ 1993
□ 1994
□ 1995
□ 1996
□ 1997
☑ 1998
☑ 1999
☑ 2000
☑ 2001
☑ 2002

alt [79]

Waterl BY / 77	Wetland BZ / 78	alt CA / 79	country CB / 80	gps CC / 81	trapID CD / 82	trap_name CE / 83	year CF / 84
2249.0	0.0	93.0	England	N55.214 E1.682	15	Newcastle	1998
2249.0	0.0	93.0	England	N55.214 E1.682	15	Newcastle	1999
2249.0	0.0	93.0	England	N55.214 E1.682	15	Newcastle	2000
2249.0	0.0	93.0	England	N55.214 E1.682	15	Newcastle	2001
2249.0	0.0	93.0	England	N55.214 E1.682	15	Newcastle	2002

□ Use reg. Exp

Figure 1. Geographical visualization of the *Myzus persicae* data set, displaying a subset of the data collected and a dialog with the details associated with the data point on the map that the user clicked on.

Results and discussion

In the experiments, we performed 10 runs of 10-fold cross-validation, using the corrected paired t-tester (Nadeau & Bengio 2003) for statistical analysis. The results for the best parameter settings for each algorithm are shown in Tables 2 and 3, showing the correlation coefficients and root mean squared errors of the observed and predicted data.

Though not significantly better than linear regression, the support vector machine (SMOreg) gave the highest correlation coefficient for predicting the Julian date, with the neural network (MLP) performing the worst. However the support vector machine performed significantly better than linear regression predicting the $\log_{10}$-transformed annual numbers with respect to both the correlation coefficient and root mean squared error.

Table 2. Comparing various regression algorithms with respect to their correlation coefficient, with LR. The symbols, ^{a} and ^{b}, indicate whether the result is significantly lower or higher, respectively, than LR at P=0.05.

Dataset	LR	M5'	MLP	GPD	SMOreg
MpeJd5thFlt	0.81±0.05	0.78±0.05	0.76±0.06^a	0.81±0.04	0.83±0.04
MpeLgTotW52	0.76±0.06	0.77±0.04	0.76±0.04	0.78±0.04	0.83±0.03^b

Table 3. Comparing various regression algorithms based on their root mean squared error, compared with LR. Symbols, ^{a} and ^{b}, indicate whether the result is significantly lower or higher, respectively, than LR at P=0.05.

Dataset	LR	M5'	MLP	GPD	SMOreg
MpeJd5thFlt	25.66±10.77	29.31±3.31	29.79±4.16	25.07±3.09	23.53±3.27
MpeLgTotW52	0.43±0.09	0.44±0.04	0.47±0.05	0.41±0.03	0.37±0.0^a

Further data mining

Inexpensive and ubiquitous sensors enable us to collect large amounts of data. However, more is not always better. Algorithms are usually sensitive to the amount and order of the data presented to them. Finding a good subset of attributes can be achieved with attribute selection. For the sake of demonstration, one question that we could ask about the country data is whether we can we identify the "country" based on the land use, land cover and climatic attributes alone? Not only do we want to have a model that explains the data very well, we also want it to be easily interpreted by humans. Decision trees, like J48 (WEKA's C4.5 implementation), are a good candidate. For this experiment, we remove the following variables from the data: *trap_name, trapID, year, long, lat, MpeJd5thFlt, MpeLgTotW52.* When built on the remaining data set, J48 generated a tree comprising 65 nodes and 33 leaves. Accuracy from a single run of 10-fold cross-validation was 99.5% (only 7 out of 1492 examples were misclassified (Figure 2).

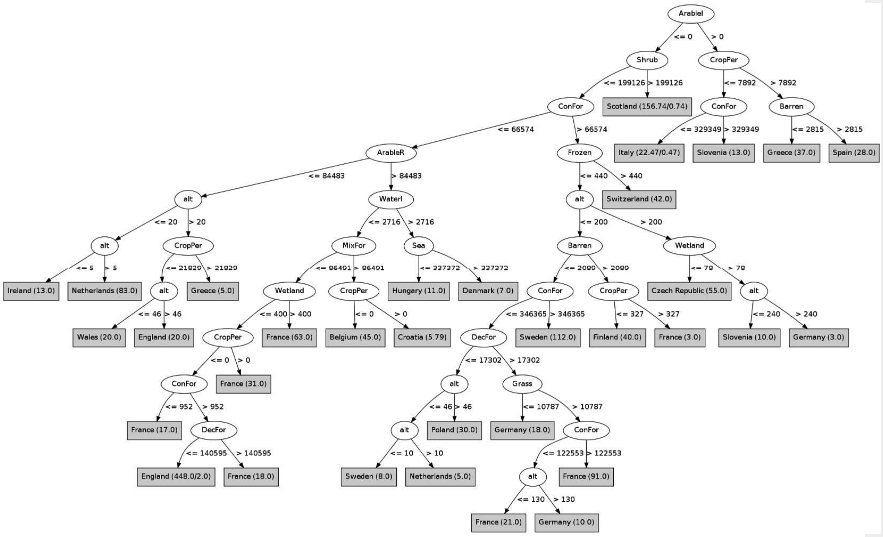


Figure 2. J48 decision tree for determining country generated from the full *Myzus persicae* data set.

Can each country be identified using fewer attributes? WEKA's attribute selection, BestFirst search algorithm using default parameters, and the WrapperSubsetEval, using J48 were applied to the data. The result was a dataset comprising only the following four landcover and landuse attributes to identify each country, the amount deciduous forest (*DecFor),* permanent crops *(CropPer),* arable farmland *(Arable),* and urban area *(Urban)*. Evaluating J48 on this reduced data set, resulted in a perfect accuracy of 100% on a single run of 10-fold cross-validation. However, this comes at the cost of larger trees. The resulting tree had 95 nodes and 48 leaves (Figure 3). Both models are very good and choosing between them depends on the objective of the model (or a combination of them). For example, a smaller tree, higher accuracy or fewer attributes. Care needs to be taken however, where higher accuracy can be misleading. High accuracy may not necessarily represent a more comprehensive model, it may mean that the model fits the noise present in the data, better. A smaller model on the other hand, can often give better generalization, though an overly aggressive pruning method can result in a too simplistic model.

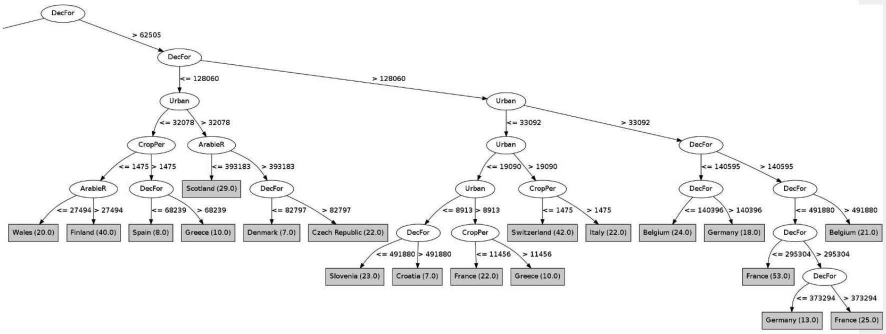


Figure 3. J48 decision tree (detail) for determining country based on variables DecFor, ArableR, CropPer and Urban from the *Myzus persicae* data set.

So far, we have analyzed the data set as a whole. What about using localized models, i.e. per country? Do they perform better or worse? We use a workflow to split the data with "MpeLgTotW52" as response variable into country subsets and evaluated them using a correlation coefficient between the observed and predicted data using the best algorithm settings obtained from the previous experiments. Since 10-fold cross-validation is used, only countries with at least 10 data entries were considered. Due to the large variation in number of entries, we expected the correlation coefficient to vary a lot. For example England had 412 data entries over site years, whereas Ireland had 12. The result is shown in Figure 4, with the bar graph showing the percentage of data entries for each country. As expected, England, France and Scotland produce the best models, since they have the highest percentage data entries. However, the Czech Republic with its five trap locations, generates a reasonably good model with only 55 data entries. This graph suggests that for most countries more data is required before reasonable individual models can be generated.

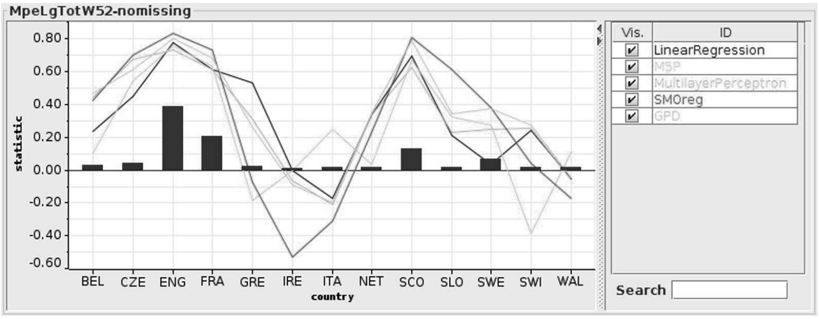


Figure 4. Per country plot of correlation coefficient obtained from single run of 10-fold cross-validation. The bar plot shows the percentage (0-1) of data entries per country from the overall data set.

Sticky plates

The previous sections discussed the development of models for crop protection. Once a model has been developed, it can be deployed. Using a data-driven workflow system, like ADAMS, the necessary collection and preprocessing steps can be specified and automated, providing the model with data in the required format and converting its output into other formats for further processing. The following sections describe briefly the development of the next generation of the Scoutbox system (http://scoutbox.nl/en/). The system, sold by Cropwatch BV is a part of the DutchSprouts BV (http://www.dutchsprouts.com/) system. Scoutbox is used for counting insects captured in traps using yellow sticky plates in greenhouses, enabling the customer to monitor trends in insect populations.

Process

The greenhouse operator places sticky plate traps in various fixed locations in the greenhouse and collects them at regular intervals by taking high-resolution 18 mega-pixel pictures using the Scoutbox system. The images are uploaded and processed in batches in the cloud. Through the Scoutbox portal, the customer can view the count history for various insects such as the greenhouse whitefly, *Trialeurodes vaporariorum*, the predatory bug, *Macrolophus caliginosus*, and western flower thrips, *Frankliniella occidentalis*), determining whether biological or chemical intervention is required.

System

In contrast to the *Myzus persicae* data set discussed earlier, no expert is required to identify and count the insects, once the models have been generated and deployed. An expert is only required during training time, to identify the insects on the sticky plates. The system locates objects, insects or otherwise, on the plate images and extracts these as cropped sub-images. These sub-images, after being presented to the expert and labelled accordingly, are then added to training sets. From these training sets of images, various features determined to be robust, through a series of experiments, are generated and used as input for data mining algorithms provided by WEKA. The ADAMS workflow system provides the image analysis functionality for locating insects on a plate, generating the features and building or applying the models. In production, the workflow also performs optical character recognition of ID tags (associating the plates with customers), generates formatted output that the Scoutbox system integrates in its data warehouse used for visualizing the insect data for the customer. The steps for locating insects and feature extraction are shared between the model building and production workflows. Figure 5 shows a schematic diagram of how the overall production system works.

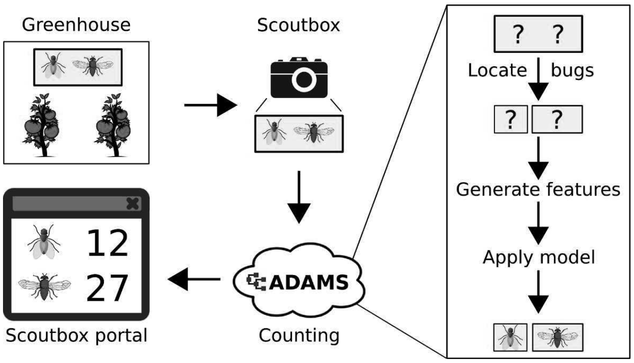


Figure 5. Schematic diagram for the data flow of the Scoutbox application, using ADAMS as its data processing back-end.

Conclusion

Data mining can be a powerful tool when analysing crop protection data provided there are sufficient amounts of it. Depending on whether the goal, is interpretability (e.g. application of a linear regression) or accuracy (e.g. using a support vector machine), different techniques may be selected. Workflows can capture the different steps involved in data collection, cleansing, transformation, prediction and output generation, documenting the complete process, therefore providing greater repeatability. As new data is created, previous experiments for evaluating models can be re-run at no extra cost to validate or invalidate previous hypotheses. The use of workflow applications also helps to demonstrate whether a research prototype, for example, a predictive model developed for publication, might be turned into a commercial production system.

Acknowledgements

Grateful thanks to Dr Richard Harrington, Head, Rothamsted Insect Survey, Agroecology Department, Rothamsted Research, Harpenden, Hertsfordshire, UK, for making the aphid data available.

References

Cocu N, Harrington R, Rounsevell MDA, Worner SP, Hullé M 2005. Geographical location, climate and land use influences on the phenology and numbers of the aphid, *Myzus persicae*, in Europe. Journal of Biogeography 32: 615-632.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1): 10-18.

Nadeau C, Bengio Y 2003. Inference for the Generalization Error. Machine Learning 52(3): 239-281.

Quinlan R J 1992. Learning with Continuous Classes. 5th Australian Joint Conference on Artificial Intelligence, Singapore. Pp. 343-348.

Rasmussen CE, Williams CKI 2006. Gaussian processes for machine learning. The MIT Press, Cambridge, MA, USA. ISBN 0-262-18253-X.

Reutemann P, Vanschoren J 2012. Scientific Workflow Management with ADAMS. Proceedings of the Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), Part II, LNCS 7524. Pp. 833–837.

Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK 1999. Improvements to the SMO Algorithm for SVM Regression. IEEE Transactions on Neural Networks 11 (5): 1188-1193.

Wang Y, Witten IH 1997. Induction of model trees for predicting continuous classes. Proceedings of the 9th European Conference on Machine Learning – Poster papers. Pp. 128-137.