Cost-senstive learning for Rule Classification: Evaluation of its applicability for Integrated Pest Management

Borja Espejo-Garcia, F.J. Lopez-Pellicer, F.J. Zarazaga-Soria

Advances Information System group (IAAA)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: borjaeg@unizar.es

Abstract

This work evaluates and compares different supervised learning algorithms using a cost-sensitive approach to find a model that classifies legal rules related to pesticides as prohibitions and permissions. The naive Bayes classifier achieves the best results and it would be applicable because it doesn't misclassify prohibitions as permissions.

Introduction

In modern agriculture, as [1] states, production is governed by norms that restrict harmful farming practices such as the use of dangerous phytosanitary products in crops. For example, the EU has set up regulations that require a mandatory use of Integrated Pest Management (IPM) in EU Member states. These regulations ask Governments to establish methods for determining whether farmers apply IPM principles.

On the other hand, as [2] explains, agriculture has entered a new era in which different information systems such as Farm Management Information Systems (FMIS) have emerged to help farmers to comply with IPM requirements and avoid fines. Although, information systems can provide law compliance assessment, they require processable rules derived from IPM norms. These rules can be encoded manually with a formal representation by an expert, but this task can be overwhelming. An alternative is the use of a rule learning system. However, missed and wrong learned rules are source of health risks whose social costs can be approximately measured by their effective economical impact (e.g. fines, operation closings, etc.). For example, if a compliance system classifies a rule that prohibits a pesticide as permission, it could cause important public health problems. It would not be the first time [3]. In order to minimise health risks in this approach, we evaluate the applicability of using natural language processing and supervised learning techniques to classify rules [4] with a cost-sensitive approach [5].

Material and Methods

For any project aiming to incorporate supervised learning, it is necessary the creation of an annotated corpus. The approach of this work is the use of the deontic logic operators *Prohibition* and *Permission* for annotating rules in such corpus. The rules annotated in this work have been extracted from the phytosanitary products register published by the Spanish Ministry of Agriculture, Food and Environment. The annotation was performed using 513 rules (177 prohibitions and 336 permissions) and then transformed into simple bag-of-words representations, where each feature is a single token.

Once the rules have been modelled as feature vectors, we apply different learning algorithms to find, if possible, a classifier that does not classify *prohibition rules* as *permission rules*. We use Naive Bayes, Random Forests and Support Vector Machines (SVM) with a linear kernel. Bayes classifiers are known for creating simple yet well performing linear models. Random Forests, which are ensembles of decision trees, are able to capture non-linearities. Finally, SVM has good performance with high dimensionality problems. In addition, since we are facing a cost-sensitive learning problem, we need to include the cost information in learning algorithms. In this work, we apply oversampling by reweighting training instances.

To evaluate the applicability of the learning algorithms, we use precision and recall. Precision is the proportion of "truly" *prohibition rules* to the total number of rules classified as *prohibition rules*. Recall is the fraction of "truly" *prohibition rules* that are effectively classified as *prohibition rules*. Our assumption is that if the compliance is fully automatic, we cannot tolerate any results different to 100% recall. Therefore, the baseline algorithm is a classifier that obtains a 100% recall by always classifying rules as *prohibition rules*.

Results

The results reported are the average of 20 runs using a 10-cross validation approach. Figure 1A shows the recall obtained with the different algorithms at different cost ratios. This experiment shows that augmenting the cost of misclassifying a prohibition, the unique algorithm that is cost-sensitive enough to obtain a 100% recall is the Bayes classifier, and, thus, it could be applicable in the IPM context. Finally, to conclude that Bayes is better than the baseline, it is necessary that it also obtains a better precision. Figure 1B shows that when the cost ratio is 1:25, Bayes classifier improves the baseline achieving a precision of 44%.

Conclusions

In this article, it has been shown that it is possible to use supervised learning techniques with a cost-sensitive approach to classify rules as *prohibition rules* and *permission rules*. Our main objective was to find an algorithm that obtains a 100% recall, i.e., prohibitions are never misclassified, and a precision higher than the baseline method. This objective has been achieved with the naïve Bayes algorithm. Moreover, as future research, we propose to increase the expressivity of the rule model and to use new approaches for cost-sensitive learning.

Acknowledge

The work of Borja Espejo-Garcia has been partially supported by a grant from the Aragon Government.

References

[1] NIKKILÄ, R, WIEBENSOHN, J, NASH, E, SEILONEN, I. and KOSKINEN, K. A service infrastructure for the representation, discovery, distribution and evaluation of agricultural production standards for automated compliance control. *Computers and Electronics in Agriculture*. 2012. Vol. 80, p. 80–88. DOI 10.1016/j.compag.2011.10.011.

[2] FOUNTAS, S., CARLI, G., SØRENSEN, C.G., TSIROPOULOS, Z., CAVALARIS, C., VATSANIDOU, A., LIAKOS, B., CANAVARI, M., WIEBENSOHN, J. and TISSERYE, B. Farm management information systems: Current situation and future perspectives. *Computers and Electronics in Agriculture* [online]. 2015. Vol. 115, p. 40–50. DOI 10.1016/j.compag.2015.05.011. Available from: http://www.sciencedirect.com/science/article/pii/S0168169915001337

[3] SHTIENBERG, Dani. Will decision-support systems be widely used for the management of plant diseases? *Annual review of phytopathology* [online]. 2013. Vol. 51, p. 1–16. DOI 10.1146/annurev-phyto-082712-102244. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23767845

[4] WYNER, Adam and PETERS, Wim. On rule extraction from regulations. *Frontiers in Artificial Intelligence and Applications*. 2011. Vol. 235, p. 113–122. DOI 10.3233/978-1-60750-981-3-113.

[5] ELKAN, Charles. The foundations of cost-sensitive learning. *IJCAI International Joint Conference on Artificial Intelligence*. 2001. P. 973–978. DOI doi=10.1.1.29.514.

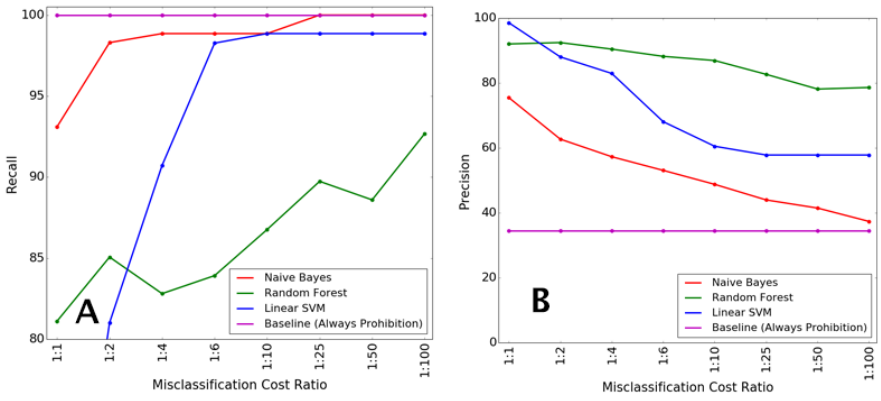

Figure 1: A) Recall comparison for every algorithm at different cost ratios. Bayes classifier is the unique classifier that obtains 100% recall B) Precision comparison. Naive Bayes obtains worse precision than Linear SVM and Random Forest, but since it is the unique that obtains perfect recall and its precision is better than the baseline, we can conclude that it is the best classifier