Computers and Electronics in Agriculture 150 (2018) 343–352

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Machine learning for automatic rule classification of agricultural regulations: A case study in Spain

Borja Espejo-Garcia[a], Jorge Martinez-Guanter[b], Manuel Pérez-Ruiz[b,*], Francisco J. Lopez-Pellicer[a], F. Javier Zarazaga-Soria[a]

[a] *Computer Science and System Engineering Department, Universidad Zaragoza, María de Luna 1, 50018 Zaragoza, Spain*
[b] *Universidad de Sevilla, Área de Ingeniería Agroforestal, Dpto. de Ingeniería Aeroespacial y Mecánica de Fluidos, Spain*

ARTICLE INFO

Keywords:
Rule extraction
Natural language processing
Smartprecision agriculture
Integrated pest management

ABSTRACT

Currently, pest management practices require modern equipment and the use of complex information, such as regulations and guidelines. The complexity of regulations is the root cause of the emergence of automated solutions for compliance assessment by translating regulations into sets of machine-processable rules that can be run by specialized modules of farm management information systems (FMIS). However, the manual translation of rules is prohibitively costly, and therefore, this translation should be carried out with the support of artificial intelligence techniques.

In this paper, we use the official Spanish phytosanitary products registry to empirically evaluate the performance of four popular machine learning algorithms in the task of correctly classifying pesticide regulations as prohibitions or obligations. Moreover, we also evaluate how to improve the performance of the algorithms in the preprocessing of the texts with natural language processing techniques. Finally, due to the specific characteristics of the texts found in pesticide regulations, resampling techniques are also evaluated. Experiments show that the combination of the machine learning algorithm Logic regression, the natural language technique part-of-speech tagging and the resampling technique Tomek links is the best performing approach, with an F_1 score of 68.8%, a precision of 84.46% and a recall of 60%. The experimental results are promising, and they show that this approach can be applied to develop a computer-aided tool for transforming textual pesticide regulations into machine-processable rules. To the best of our knowledge, this is the first study that evaluates the use of artificial intelligence methods for the automatic translation of agricultural regulations into machine-processable representations.

1. Introduction

In modern agriculture, production is governed by a variety of standards that restrict potentially harmful farming practices (Nikkilä et al., 2012). For example, different regulations and programmes, such as Integrated Pest Management (IPM), have been developed to control the use of phytosanitary products and prevent unauthorized uses (Lozano et al., 2010). IPM can be implemented as a module of a farm management information system (FMIS), which collects, exchanges and stores a large amount of exploitation data and provides decision support for tailoring a farm operation to the specific demands of stakeholders (Sørensen et al., 2010). Fountas et al. (2015) extend this model by defining a complex information ecosystem established around the farm machinery named Farm Machinery Management Information System (FMMIS), based on the Soft System Methodology (SSM). The interrelations presented in this extended model, such as GNSS positioning data, real-time crop and soil data generated by airborne or terrestrial sensors, and input consumption and inventory management databases, have been increasingly studied and adopted by farmers (Miller et al., 2017). Making farmer's decision process easier is essential.

One of the FMMIS challenges is active support during the decision-making process, which could ensure that agricultural tasks such as fertilization and spraying are conducted according to safety and quality standards. To achieve this goal, it is necessary to translate standards and regulations into a machine-readable representation, such as formal rules, that can be executed within an FMIS.

According to Nash et al. (2011) and our own experience working with the official Spanish phytosanitary products registry, agricultural regulations consist of rules that can be mainly classified as prohibitions

* Corresponding author at: Área de Ingeniería Agroforestal, Dpto. de Ingeniería Aeroespacial y Mecánica de Fluidos, Ctra. Utrera, km 1, Sevilla 41013, Spain.
E-mail address: manuelperez@us.es (M. Pérez-Ruiz).

https://doi.org/10.1016/j.compag.2018.05.007
Received 1 March 2018; Received in revised form 4 May 2018; Accepted 5 May 2018

and obligations. Thus, each of these rules can be evaluated as true or false, with the conclusion of compliance or violation of the regulation.

The manual translation of regulations into machine-processable representations is prohibitively costly in terms of time, labour and knowledge (Wyner and Governatori, 2013). Another barrier to the actual situation encompassing European farming is that most of the data and information are unstructured, fragmented and difficult to use (Fountas et al., 2015).

To avoid these bottlenecks, techniques related to artificial intelligence, such as information retrieval, natural language processing (NLP) and machine learning (ML), can be used to identify syntactical patterns in the rules and partially automate the translation of regulations into formal rules that can then be provided to the FMIS. Moreover, in recent years, some promising results have been obtained in extracting rules from regulations in several domains (e.g., Soria et al., 2005; Wyner and Peters, 2011; Maat and Winkels, 2008). In the agricultural domain, these techniques must prove that they are highly accurate because non-compliance caused by an extraction error may carry a considerable economic penalty (Davies and Hodge, 2006). Automatic rule extraction from regulations is a complex process that requires different components. One of these components should be a rule classifier that allows regulations to be categorized as prohibitions or obligations (Fig. 1). This step is critical because an error implies that the meaning of the rule is inverted. For example, a rule such as "*Do not apply to crops with fruits that must be preserved*" could be interpreted as "*Apply to crops with fruits that must be preserved*". Moreover, this classification could facilitate the modelling conditions and rule constraints that represent the meaning of rules and retain consistency with the original text. There are different ways to build rule classifiers, but the state-of-the-art approach includes the use of ML algorithms. Moreover, these algorithms are often enriched with linguistic knowledge that is automatically extracted by using NLP techniques and improved by using preprocessing techniques such as resampling.

This work evaluates the applicability of NLP, resampling and ML techniques for building a rule classifier that can automatically discern between prohibitions and obligations in the agricultural domain using documents from the official Spanish phytosanitary product registry. We have preprocessed these documents to extract only the parts of the text that represent the rules. Then, we have manually annotated them to create a gold corpus in which the ML algorithms will find the patterns that allow the distinction between prohibitions and obligations. This gold corpus will also be used as a benchmark to evaluate the performance of the different techniques evaluated in this paper. It is important to note that we have created our own corpus because, as far as we know, there is no available gold corpus focused on phytosanitary regulations, in contrast to other research domains such as spam classification or news categorization. In this paper, we provide insights into the possibilities and limitations of existing ML, resampling and NLP techniques for usage in agriculture to support the development of decision support systems and the FMIS. Moreover, the objective of the approach presented in this paper is to provide a basis for the future automatic extraction of rules and their spatiotemporal components. As noted by Nikkilä et al. (2012), we believe that the fully automated translation of regulations is not currently feasible, but building knowledge repositories and software components that gradually solve the rule translation problems will benefit future studies.

This article is structured as follows: Section 2 presents the materials and methods and the complexity of the FMIS, and the developed methodology is detailed. Section 3 shows the results of the evaluation of the techniques analysed, and Section 4 presents a discussion of its implementation in FMIS's. Finally, Section 5 presents the conclusions and future directions for the integration of these techniques in modern FMIS's.

2. Materials and methods

2.1. Commercial FMIS structure and data sharing enhancements

In modern agrisystems, many devices, including tractors, tractor implements, field sensors, and airborne devices, are used on farms. The information generated and required by these devices must be understandable to optimize collaboration efforts. To simplify the above-mentioned interconnection of different farm elements and provide a unified data platform, the commercial solution Agroplanning was created. Agroplanning is a modular cloud-based FMIS that treats the tractor as a centralized connected platform for data generation and reception. The aim of the system is to incorporate the tractor-centric approach defined by Fountas et al. (2015) and to equip agricultural service companies, farmers, cooperatives and machinery manufacturers with the tools to generate the first advanced precision farming services,

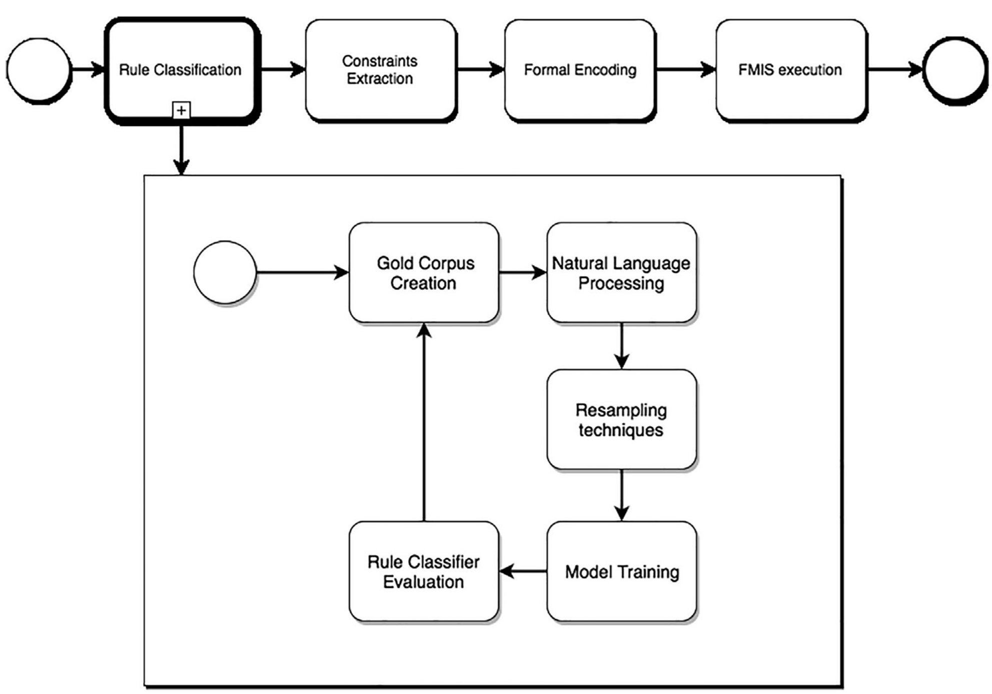

Fig. 1. Processes for translating a rule into a machine-readable format. Rule classifier development is an interactive process. If a regulation changes, it may be necessary to retrain the classifier with new data. Developing a rule classifier is a complex task that requires the combination of NLP techniques, resampling methods and model training.