Engineering Applications of Artificial Intelligence 22 (2009) 579–592

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Using Bayesian networks with rule extraction to infer the risk of weed infestation in a corn-crop

Gláucia M. Bressan [a], Vilma A. Oliveira [a,*], Estevam R. Hruschka Jr.[b], Maria C. Nicoletti [b]

[a] Universidade de São Paulo, Departamento de Engenharia Elétrica, 13566-590 São Carlos, SP, Brazil
[b] Universidade Federal de São Carlos, Departamento de Computação, 13565-905 São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 30 November 2007
Received in revised form
20 January 2009
Accepted 24 March 2009
Available online 14 May 2009

Keywords:
Bayesian network
Naïve Bayes
Rule extraction
Weed infestation
Kriging

ABSTRACT

This paper describes the modeling of a weed infestation risk inference system that implements a collaborative inference scheme based on rules extracted from two Bayesian network classifiers. The first Bayesian classifier infers a categorical variable value for the weed–crop competitiveness using as input categorical variables for the total density of weeds and corresponding proportions of narrow and broad-leaved weeds. The inferred categorical variable values for the weed–crop competitiveness along with three other categorical variables extracted from estimated maps for the weed seed production and weed coverage are then used as input for a second Bayesian network classifier to infer categorical variables values for the risk of infestation. Weed biomass and yield loss data samples are used to learn the probability relationship among the nodes of the first and second Bayesian classifiers in a supervised fashion, respectively. For comparison purposes, two types of Bayesian network structures are considered, namely an expert-based Bayesian classifier and a naïve Bayes classifier. The inference system focused on the knowledge interpretation by translating a Bayesian classifier into a set of classification rules. The results obtained for the risk inference in a corn-crop field are presented and discussed.

1. Introduction

Agricultural procedures may modify the ecological balance of a field due to the tilling procedures growers use to prepare the land, quite often leading to a population explosion or infestation of some inconvenient plants commonly known as weeds. Weed control is a fundamental part of all crop production systems. Yield reductions due to weeds are commonly known obstacle in harvest operations as they lower crop quality by competing with the crop for limited resources, such as water, nutrients, light, etc. Oerke et al. (1994) estimated that a 10% loss of worldwide agricultural production might be a consequence of weed activity.

In general, the main components of weed management systems are herbicides. Usually, herbicides are uniformly spread over the entire field aiming at weed control. A uniform application rate is often based on a visual evaluation of the weed density, with no procedure used to evaluate the risks associated with under and over spraying (Faechner et al., 2002). However, weed infestation does not occur over the entire field and the amount of herbicides could be reduced by spraying only over the weed patches (Wallinga et al., 1998; Jurado-Expósito et al., 2004). The prediction of weed dispersion can be efficiently used in preventing infestations by applying herbicides only in specific regions (Jurado-Expósito et al., 2003; Faechner et al., 2002). Reducing the quantity of herbicides potentially reduces herbicide residues in water, food crops and in the environment, and it may prevent the development of weed resistance (Aitkenhead et al., 2003).

In the literature, a considerable diversity of weed management decision models can be found. There are many different approaches, ranging from empirical functions to mechanistic simulation models. As surveyed by Wilkerson et al. (2002), some of the models are too simple as they do not include all factors that can influence weed competition or other issues farmers consider when deciding how to manage weeds. Other models can be excessively complex given that many users might find difficulty in obtaining the needed information or do not have the required equipment for acquiring the data. According to Wilkerson et al. (2002), weed management decision models must be built and evaluated from three perspectives: biological accuracy, quality of recommendations and ease of use. In addition, another important issue to be taken into account when building weed management systems is related to the interpretation of the model. The latter is of particular interest in the experiments conducted in this paper.

There are few formalisms that can be used to model weed infestation in a crop field. Primot et al. (2006) developed 20 simple models (five are linear regression models and the other 15

* Corresponding author. Tel.: +55 16 33739336; fax: +55 16 33739372.
E-mail address: vilmao@sel.eesc.usp.br (V.A. Oliveira).

0952-1976/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2009.03.006

are logistic regression models). The models were evaluated for their ability to discriminate the fields with a high level of weed infestation from the fields with a low level of infestation—the parameters of the 20 models were estimated using 3 years of experimental data. The models can be used to help farmers decide what type of weed control (chemical, mechanical or biological) to use.

The risk of weed infested crop can be inferred from the mathematical modeling of the weed behavior, based on experimental data. Dynamic models for weed seed populations describe the population size at life-cycle t as a function of the population size at life-cycle $t-1$ using difference (Sakai, 2001; Cousens and Mortimer, 1995). The dynamic models indicate that infestation is not only dependent upon the weed density but also on the competitiveness of the weed species (Park et al., 2003; Firbank and Watkinson, 1985; Kropff and Spitters, 1991). More recently, competitive indexes and weed ranking were used to quantify the weed competitiveness in a soybean field (Hock et al., 2006). Although purely mathematical models can be used for modeling the weed risk of infestation, with good performance, as described in several of the previous references, most of them lack flexibility and more important, lack interpretability—they work as 'black boxes' where the user feeds a few values and the system outputs a diagnosis.

A particular class of models is based on probability. Of special interest in this paper is the class of Bayesian networks (BN) models, which are based on the probability that a given set of measurements define objects as belonging to a certain class. In the literature, Bayesian based methods have already been used for modeling similar problems (Hughes and Madden, 2003; Smith and Blackshaw, 2002; Banerjee et al., 2005). Particularly, Hughes and Madden (2003) proposed a risk assessment methodology to identify which exotic plant species, among those presented for import, are a threat (to agricultural and ecological systems) and which are not. Bayesian theory has also been employed in the agriculture domain as the basis for developing classification systems, as described in Granitto et al. (2002). In their work, the performance of a naïve Bayes classifier (BC) is used as the selection criterion for identifying a nearly optimal set of 12 seed characteristics further used as classification parameters, such as coloration, morphological and textural features. Considering the seed identification problem, the work described in Granitto et al. (2005) compared naïve Bayes classifier performance to an artificial neural network (NN) based classifier. In this particular experiment the naïve Bayes classifier with an adequately selected set of classification features outperformed the NN based classifier. Similar result was also obtained in Marchant and Onyango (2003) but with a Bayesian classifier and a multilayer feed-forward neural network in a task for discriminating plants, weeds, and soil in color images.

The main goal of this paper is to propose and describe the use of Bayesian network methods to infer the risk of weed infestation in a corn-crop as well as to present and discuss the results obtained in a real application domain based on empirical data. The procedure is implemented as a collaborative system that integrates two classification tasks. The first uses a Bayesian network to infer the competitiveness of weeds expressed by their biomass, using as input the total density of weeds, and corresponding narrow and broad-leaved proportions. The second task assesses the risk of infestation, expressed by the yield loss, using as input the previous inferred competitiveness, as well as features extracted from the weed seed density, weed coverage and weed seed patches. The three last variables are estimated with a geostatistics method called kriging (Brooker, 1979; Isaaks and Srivastana, 1989) and image objects (Gonzalez and Woods, 2002) from weed seed density and weed coverage data samples.

In addition, the paper also presents the translation of the induced Bayesian networks into a set of classification rules, aiming at a more comprehensible knowledge representation. As mentioned before, this is an important aspect of a knowledge based system construction, since it provides the system credibility, a quality that other types of representation lack. Therefore, the main idea of the conducted experiments is not to show that the translation method is better than traditional classifiers (as C4.5, for instance) or rule extraction methods. The claim is that it is possible to take advantage of both the causal knowledge representation (which can be adequately represented in a BN or BC) and high accuracy of a Bayesian classifier to have a set of classification rules (extracted from the BC) as a knowledge base.

For both classification tasks implemented by the collaborative system, two different Bayesian network structures are used for comparison purposes. One is induced by the naïve Bayes algorithm (Duda and Hart, 1973) using empirical data and the other, an unrestricted Bayesian network, is designed and refined by an expert using the same empirical data. The networks in this paper are referred to as naïve Bayes and expert-based networks, respectively. Due to their different architectures, the two Bayesian networks have different performances, depending on the available information. A set of probabilistic classification rules is then extracted from each of the Bayesian networks using a Markov-based strategy proposed in Hruschka et al. (2008). To reduce the number of rules where the Markov-based strategy does not remove categorical variables, a pruning strategy is proposed. The pruning strategy is mainly motivated by the fact that no extra computation effort is needed. The pruning can be done by considering only the rules having estimated probability higher than a predefined threshold. This paper is an extended and revised version of two earlier conference papers namely Bressan et al. (2007a, b).

The remaining of this paper is organized as follows. Section 2 describes the basics of Bayesian networks and naïve Bayes classifiers and discusses the importance of improving their understandability. Section 3 focuses on two important issues: the approach used to collect and to interpolate empirical data, and the construction of the collaborative system that integrates two Bayesian classifiers. Section 4 presents the results of the collaborative system, focusing on the results of the individual classifiers, that is, the Bayesian network and the naïve Bayes classifiers. Finally, Section 5 presents some concluding remarks and highlights the next steps for this research work.

2. Basics of Bayesian networks, Markov blanket and classification rules

As pointed out in Heckerman et al. (2000), Bayesian networks and Bayesian classifiers are usually employed in data mining tasks mainly because they (i) may deal with incomplete data sets straightforwardly; (ii) can learn causal relationships; (iii) may combine prior knowledge with patterns learnt from data and (iv) can help to avoid overfitting.

A Bayesian network can be viewed as a form of probabilistic graphical model used for knowledge representation and reasoning about data domains. Instead of encoding a joint probability distribution over a set of random variables, as usually done by a Bayesian network, a Bayesian classifier usually aims to correctly predict the value of a discrete class variable given the value of a vector of features (predictors). Since Bayesian classifiers are a particular type of Bayesian networks the concepts and results described in this section are valid for both.

A Bayesian network consists of two components—a network structure, which is a directed acyclic graph, and a set of