Computers and Electronics in Agriculture 99 (2013) 135–145

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Imperfect knowledge and data-based approach to model a complex agronomic feature – Application to vine vigor

Cécile Coulon-Leroy [a,*], Brigitte Charnomordic [b], Marie Thiollet-Scholtus [a], Serge Guillaume [c]

[a] INRA UE1117, UMT Vinitera, 49071 Beaucouzé, France
[b] INRA Supagro, UMR MISTEA, 34060 Montpellier, France
[c] Irstea, UMR ITAP, 34196 Montpellier, France

A R T I C L E I N F O

Article history:
Received 31 January 2013
Received in revised form 4 September 2013
Accepted 17 September 2013

Keywords:
Fuzzy logic
Knowledge imprecision
Hidden variables
Automatic learning
Data selection
Data inconsistency

A B S T R A C T

Vine vigor, a key agronomic parameter, depends on environmental factors but also on agricultural practices. The goal of this paper is to model vine vigor level according to the most influential variables.

The approach was based upon a collected dataset in a French vineyard in the middle Loire valley and the available expert knowledge. The input features were related to soil, rootstock and inter-crop management, the output was an expert assessment of vine plot vigor. The approach included a data selection step, which was needed because of data imperfection and incompleteness. Usually implicit in the literature, data selection was carried out with explicit criteria. Then a fuzzy model was designed from the selected data. Owing to the fuzzy model interpretability, its structure and behavior were analyzed.

Results showed that, despite the data imperfection, the approach was able to select data that yielded an informative model. Well-known relationships were identified, and some elements of new or controversial knowledge were discussed.

This is an important step towards the design of a decision support tool aiming to adapt the agricultural practices to the environment in order to get a given vigor level.

1. Introduction

In modern agriculture, an important issue is to optimize the agricultural practices according to environmental factors, in order to reach a given yield level and product quality. Models can be used as support for decision making (Affholder et al., 2012).

In general, agricultural systems are complex systems; this is the case of vine growing. Vegetative vine development, called ‘vigor’, takes into account the rhythm and the intensity of the vine shoot growth (Carbonneau et al., 2007). Empirically, in relative terms, vine vigor level of a plot is well known as being stable over the years (Johnson, 2003; Kazmierski et al., 2011). It is highly influenced by environmental factors, such as soil or climate, but can also be modified by agricultural practices (choice of rootstock, inter-row management, pruning type, among others). Vine vigor is a key parameter to control the balance between vegetative growth and productivity, which influences berry composition and then wine characteristics (Bramley et al., 2011; Kliewer and Dokoozlian, 2005).

Some complex mathematical models are available for vine development. These models work at a very large scale and for contrasting environmental conditions (Garcia de Cortazar Atauri, 2007; Valdes-Gomez et al., 2009). Some of them were designed for decision support with respect to very specific problems as the salinity in Australia (Walker et al., 2005). Some other models were not validated under various field conditions (Nendel and Kersebaum, 2004). For complex systems, it is difficult to design formal mathematical models. An alternative approach consists in deriving empirical statistical models from experiments.

However, for perennial crops such as grapevine, full experimental designs to test a large number of factors in interaction are very difficult to implement. On-going research consists, in most cases, in experimentally quantifying the impact of one variable on vine development while the other variables are being fixed *e.g.* (Bavaresco et al., 2008). Even if, at vineyard scale, interactions between variables involved in the agricultural system are empirically observed by winegrowers, these observations are not sufficient to analyze the functioning of the agronomical system. A special case of interesting interactions is the simultaneous impact of some environmental factors and agricultural practices. Some interactions between variables have been highlighted for vine vigor *e.g.* interactions between cover crop and water supply (Celette et al., 2005), or between cover crop and rootstock (Barbeau et al., 2006; Hatch et al., 2011). To identify these interactions is an important step toward a decision support system to adapt agricultural practices to the environment. However, vine vigor is difficult to model

* Corresponding author. Present address: LUNAM Université, Groupe ESA, UPSP GRAPPE, 55 rue Rabelais, BP30748, 49007 Angers, France. Tel.: +33 241235555.
E-mail address: c.coulon@groupe-esa.com (C. Coulon-Leroy).

0168-1699/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compag.2013.09.010

from experiments, essentially for two reasons. Firstly, the collected data are tainted with uncertainty; the features can suffer from imprecision, especially when they are assessed by human beings. Secondly, the data set is likely to be incomplete, because the agronomical system has some hidden features that are unknown or hard to assess. Due to these hidden features, the data base will probably include conflicting data: similar recorded combinations of input features may have contradictory output assessment.

Therefore it is important to include a data selection step in the modeling approach. In the literature, that step is often implicit and not described. In this paper, a selection method with explicit criteria is proposed.

Once the data are selected, various learning methods can be used to produce a model to study interactions between variables. They include artificial intelligence or statistical techniques. Both can deal with some kinds of data imperfection and both have been used in environmental modeling (Chen et al., 2008).

Common choices include classical linear models (LM) and decision trees (DT), or for more recent developments, Bayesian networks (BN). These statistical models are efficient in a wide range of situations, and often yield a confidence interval, since they are based on probability theory. However, they may be difficult to interpret or to use in cases where data imperfection and uncertainty is prevalent. For instance, LM being based on a least squared error fit, the intercept can be out of range. The coefficients associated to the terms are related to the overall influence of the corresponding feature or interaction, but do not allow an in-depth analysis of the model behavior at a more local level.

DT are easy to interpret, and have proven very useful for discriminant feature selection but this is not the main objective here. BN can incorporate expert knowledge and yield a graphical model easy to read, provided the number of nodes is not too high. They have been used for diagnosis purposes (Sicard et al., 2011). There are also some clear limitations to BN with respect to the proposed application. It may be difficult for experts to express their knowledge in terms of probability distributions. BN also have a limited ability to deal with continuous data, and discretization assumptions can significantly impact the results. Structure learning of a BN is still an open challenge, and the learning methods have a high complexity. Furthermore, as all statistical methods, they require a large amount of data to produce significant results, which is not always possible to get.

Fuzzy logic and inference systems (FIS) are part of artificial intelligence techniques. In FIS, fuzzy logic is used as an interface between the linguistic space, the one of human reasoning, and the space of numerical computation. FIS handle linguistic concepts, *e.g.* High or Low, implemented using fuzzy sets. Data imprecision is taken into account thanks to a progressive transition between the qualitative labels used for input or output variables. Fuzzy models are able to represent imprecise or approximate relationships that are difficult to describe in precise mathematical models. Historically, FIS were designed from expert knowledge (Mamdani and Assilian, 1975). This approach is limited to small systems and may give poor accurate results. Specific learning algorithms for FIS have then been proposed by Guillaume and Charnomordic (2012a) and by Guillaume and Magdalena (2006).

Fuzzy logic based models are interpretable, under a few restrictions (Guillaume and Charnomordic, 2011), this being particularly important for decision support (Alonso and Magdalena, 2011).

Fuzzy modeling was used in a previous work to predict the vine vigor imparted by the environment (Coulon-Leroy et al., 2012). The objective of the present paper is to propose a more ambitious work using fuzzy modeling to study the interactions between environmental factors, agricultural practices and vine vigor. The approach pays a particular attention to data selection, which is a critical step in supervised learning; even it is usually not explicitly dealt with in

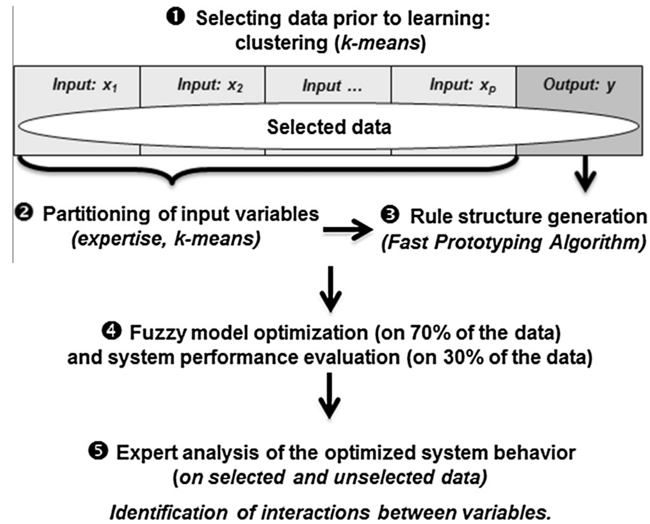

Fig. 1. Overall modeling procedure: data selection, FIS design and tuning, result interpretation and analysis.

the literature. It attempts to make the best of domain expertise and of available field data, though they are incomplete, in order to design an interpretable model. The interpretability makes it possible to analyze the system behavior and to evaluate interactions between variables.

2. Material and methods

The procedure that was used included five steps, namely:

– To describe the case study with its input and output variables of (Section 2.1).
– To select data used prior to the automatic learning (Section 2.2) by clustering (Section 2.1.1), generating sub-clusters (Section 2.1.2) and selecting consistent sub-clusters (Section 2.1.3).
– To build the fuzzy model (Section 2.3) by partitioning input variables according to data and expertise (Section 2.3.1) and generating 'if-then' rules from data (Section 2.3.2).
– To optimize the fuzzy model and to evaluate the system performance (Section 2.4).
– To analyze the optimized system and the interaction between variables (Section 2.5).

The overall procedure is summarized in Fig. 1. The multidimensional data are denoted by $(x_1, x_2 \ldots x_p, y)$ where x_i $(i = 1, \ldots, p)$ are the input variables, and y is the output variable. In the following, the output variable is a categorical variable with a given number of ordered levels.

All of the developments described in the present work are accessible using the R software (R Development Core Team, 2008) and the FisPro toolbox (Guillaume and Charnomordic, 2012a). R[1] is a free software environment for statistical computing and graphics. FisPro[2] is an open source software that corresponds to ten years of research and software development on the theme of learning interpretable fuzzy inference systems from data. It has been used in the fields of agriculture and environment (Colin et al., 2011; Coulon-Leroy et al., 2012; Rajaram and Das, 2010; Tremblay et al., 2010).

[1] http://www.r-project.org
[2] http://www7.inra.fr/mia/M/fispro/