Computers and Electronics in Agriculture 151 (2018) 61–69

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Review

Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review

Anna Chlingaryan[a,*], Salah Sukkarieh[a], Brett Whelan[b]

[a] Australian Centre for Field Robotics, Dept. of Aerospace, Mechanical & Mechatronic Engineering, The University of Sydney, NSW 2006, Australia
[b] Centre for Carbon, Water and Food, School of Life and Environmental Sciences, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Vegetation indices
Features extraction
Predictive modelling
Decision making
Information fusion

ABSTRACT

Accurate yield estimation and optimised nitrogen management is essential in agriculture. Remote sensing (RS) systems are being more widely used in building decision support tools for contemporary farming systems to improve yield production and nitrogen management while reducing operating costs and environmental impact. However, RS based approaches require processing of enormous amounts of remotely sensed data from different platforms and, therefore, greater attention is currently being devoted to machine learning (ML) methods. This is due to the capability of machine learning based systems to process a large number of inputs and handle non-linear tasks. This paper discusses research developments conducted within the last 15 years on machine learning based techniques for accurate crop yield prediction and nitrogen status estimation. The paper concludes that the rapid advances in sensing technologies and ML techniques will provide cost-effective and comprehensive solutions for better crop and environment state estimation and decision making. More targeted application of the sensor platforms and ML techniques, the fusion of different sensor modalities and expert knowledge, and the development of hybrid systems combining different ML and signal processing techniques are all likely to be part of precision agriculture (PA) in the near future.

1. Introduction

Improving crop yield production and quality while reducing operating costs and environmental pollution is a key goal in precision agriculture (PA). The potential growth and yield depends on many different production attributes such as the weather, soil properties, topography, irrigation and fertilizer management. The need for timely and accurate sensing of these inputs for large agricultural fields has led to increased adoption of remote and proximal sensing technologies (Campbell and Wynne, 2011) in PA (Curran, 1987). These sensing techniques provide acquisition of spectral, spatial and temporal information about the objects via ground-based vehicles, aircraft, satellites and handheld radiometers.

Remote sensing, such as satellite and airborne multi-spectral scanning, photography and video, enables precision weed management through the generation of timely and accurate weed maps (Lamb and Brown, 2001). Thermal remote sensing via airborne thermal imagery has the potential to identify spatial variations in crop water status (Tilling et al., 2006), which can enable improvements in the water management in irrigated cropping systems. Nowadays, this remote sensing technique is widely used in different crop species such as wheat (Gontia and Tiwari, 2008), maize (Taghvaeian et al., 2012), trees (Bellvert et al., 2016) and vineyards (Gutiérrez et al., 2018). At the same time many ground-based platforms have been developed and aimed at different PA tasks such as mapping of soil properties (Barnes et al., 2003), estimating evapotranspiration and drought stress (Maes and Steppe, 2012), weed mapping (Sui et al., 2008) and assessing crop water and nitrogen status (El-Shikha et al., 2007; Govender et al., 2009).

Remote sensing at visible and near-infrared wavelengths (vis-NIR) has been used to devise many spectral indices for estimating different vegetation properties. This includes the amount of chlorophylls and other photosynthetic/photoprotective pigments and the leaf area index (LAI) (Barati et al., 2011; Gao, 1996; Haboudane et al., 2004; Huete, 1988; Qi et al., 1994; Sims and Gamon, 2002; Viña et al., 2011; Zarco-Tejada et al., 2012). More than 100 vegetation indices along with their applicability, representativeness, environment and implementation precision have recently been reviewed by Xue and Su (2017). They concluded that for real-world applications the use of any existing vegetation indices requires careful consideration of the strengths and shortcomings of those indices and the specific environment where they will be applied. Making crop yield predictions using remotely sensed

* Corresponding author.
E-mail address: a.chlingaryan@acfr.usyd.edu.au (A. Chlingaryan).

https://doi.org/10.1016/j.compag.2018.05.012
Received 24 November 2017; Received in revised form 7 May 2018; Accepted 8 May 2018
0168-1699/ Published by Elsevier B.V.

vegetation indices has been attempted by Panda et al. (2010), Jaafar and Ahmad (2015) and many others.

Utilization of wireless sensors and actuators in PA, as well as algorithms for wireless sensor network data integration is now advancing (Zhou et al., 2012). Aqeel-ur-Rehman et al. (2014) presented a review on wireless sensor network technology and their applications in different aspects of agriculture and reported on existing system frameworks in PA.

Nitrogen (N) is considered by growers as a major mineral nutrient for plant growth and development because it is directly related to the photosynthetic process (Andrews et al., 2013). At the same time N has a high environmental and economic impact. Hence, the optimization of N fertilization for different crops has become a subject of many spectrometric studies (Cao et al., 2017; Chen et al., 2008; Goron et al., 2017; Lukina et al., 2001; MacKerron et al., 1993; Maresma et al., 2016; Quemada et al., 2014; Raun et al., 2005; Schepers and Raun, 2008).

The estimation of the plant N status can be divided into two main types: destructive and non-destructive. The most common method of destructive measurement is a chemical analysis which is associated with the Kjeldahl technique and is laborious, lengthy and costly (Jones and Moseley, 1993; Vigneau et al., 2011). Optical remote sensing of the plant N status is a non-destructive method based on canopy reflectance in the visible–NIR wavelengths (400–900 nm). This measurement is completed in-situ, lowering the number of field samples required and thus reducing the time and financial cost of field sample collection, preparation and laboratory analysis. Many studies have been dedicated to non-destructive measurement of the N status inferring in plants via remote sensing technology (Apostol et al., 2007; Lamb et al., 2002; Reyniers and Vrindts, 2006; Scharf et al., 2002; Tremblay et al., 2012) and spectral indices indicative of the plant N status have been derived from hyperspectral data (Chen et al., 2010; Tian et al., 2014; Yao et al., 2010).

Adoption of remote sensing in geology (Gupta, 2003), forestry (Holmgren and Thuresson, 1998; Hultquist et al., 2014), hydrology (Engman and Gurney, 1991), agriculture (Seelan et al., 2003) and other domains has led to the collection of significant volumes of data. The volume is continuously growing and it is beyond human ability to personally integrate, analyse and make the best informed decisions from the information. This is particularly the case when the data is not homogeneous, i.e. is sensed by sensors with different spatial, temporal and spectral modalities. Machine Learning (ML) is an emerging technology that can aid in the discovery of rules and patterns in large sets of data (Du and Jeffrey, 2007).

Crop yield prediction and N status estimation are considered together here because of the direct linkage in fertiliser management decisions. Crop yield goals are routinely utilised for calculating N requirements, both pre- and in-season. For devising potential site-specific management plans for N fertiliser, especially in-season, an estimation of both would be ideal. The aim of this review is to show the capability of different ML techniques to effectively handle these different but closely related tasks. A review is presented of recent studies in the area of crop yield prediction and N status estimation, which incorporate different ML techniques. It also covers comparative studies of different ML techniques as they are applied to the same task in PA. Some technical aspects of the ML techniques used in the reviewed studies are discussed.

2. Machine learning techniques

One of the main advantages of ML techniques is that they are capable of autonomously solving large non-linear problems using datasets from multiple (potentially interconnected) sources. Some ML techniques, such as Gaussian Processes (GPs) (Bishop, 2006; Rasmussen and Williams, 2005), Dirichlet Processes (DP) (Ferguson, 1973) and Indian Buffet Process (IBP) (Griffiths and Ghahramani, 2011) are probabilistic and enable consideration of sensor noise while conducting probabilistic fusion of information from different sensors (Castaldi et al., 2016; Dalponte et al., 2012; Pohl and Van Genderen, 1998) and providing confidence intervals for the predictions. ML enables better decision making and informed actions in real-world scenarios without (or with minimal) human intervention. ML provides a powerful and flexible framework for not only data-driven decision making but also for incorporation of expert knowledge into the system. These are some of the key characteristics of the ML techniques that make them widely used in many domains, and highly applicable to PA.

The major aim of PA in cropping systems is to provide information that will enable better decisions to be made on management across space and time (Whelan and Taylor, 2013). Specifically, information on variation in plant health and physiology, nutrient status or pest/disease burden may allow different treatments or treatment intensity to be applied to specific areas of crop. The practical identification and segregation of areas is commonly achieved by dividing a large field area into smaller management zones with identified requirements for different treatments. Conventionally, such delineation is based on maps of the crop field variability derived from soil and yield measurements. Nawar et al. (2017) provided a comprehensive review on management zone delineation approaches for PA applications. They illustrated how recent developments in sensing technologies, geostatistical analysis, data fusion and interpolation techniques have improved precision and reliability of management zone delineation, making it a viable strategy in commercial agriculture. They also compared traditional with advanced sensing technologies for delineating management zones.

Recently, Pantazi et al. (2015) demonstrated that the combination of data fusion modelling and clustering methods was able to improve the quality of management zone delineation. Specifically, they compared k-means clustering with the Self-Organizing Map (SOM) for delineating management zones maps for variable-rate N application. Furthermore, a hybrid SOM algorithm in combination with k-means was compared with k-means in terms of cluster separation and management zone formation based on data fusion of NDVI and soil parameters.

As reviewed by Behmann et al. (2015), ML techniques have been widely used for the early and accurate detection of biotic stress in crop, specifically, for detection of weeds, plant diseases and insect pests. Mehra et al. (2016) used ML techniques such as Artificial Neural Networks (ANNs), categorical and regression trees and Random Forests (RFs) to approach the problem of predicting the pre-planting risk of Stagonospora nodorum blotch (SNB) in winter wheat. They developed risk assessment models that could be useful in making disease management decisions prior to planting of the wheat crop. Also Tellaeche et al. (2008) showed that cost savings and reduced pollution could be achieved by a Bayesian framework based automatic decision making process for detecting weeds in corn crops.

Machine learning techniques applied to hyperspectral imaging data can be used to reveal physiological and structural characteristics in plants and enable tracking physiological dynamics due to environmental effects (Wahabzada et al., 2016). Goldstein et al. (2017) demonstrated that field data, such as soil moisture, weather, irrigation characteristics, and resulting yield could be fused via ML techniques to provide automated recommendations for irrigation. Gutiérrez et al. (2018) have used thermal imaging and a combination of two ML techniques (Rotation Forests and Decision Trees) to develop a new methodology for the on-the-go assessment of vineyard water status with potential for irrigation decision making.

Machine learning techniques can be used in field spectroscopy for offline and online prediction of the soil parameters studied in the field (Morellos et al., 2016). They can work not only with variables such as derived spectral indices, but also with the entire spectral response trace (Wittenberghe et al., 2014). Spectral indices depend on a small number of available spectral bands and therefore don't use the entire information conveyed by the spectral trace. Thus, there is always the question: which vegetation index or suite of vegetation indices is better for the given task? (Panda et al., 2010). ML techniques, such as Neural