Computers and Electronics in Agriculture 127 (2016) 467–474

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Soybean varieties portfolio optimisation based on yield prediction

Oskar Marko *, Sanja Brdar, Marko Panic, Predrag Lugonja, Vladimir Crnojevic

BioSense Institute, Dr Zorana Djindjica 1, 21000 Novi Sad, Serbia

ARTICLE INFO

Article history:
Received 31 March 2016
Received in revised form 5 July 2016
Accepted 9 July 2016

Keywords:
Yield prediction
Seed selection
Weighted histograms
Portfolio optimization
Convex optimization

ABSTRACT

One of the biggest problems in agriculture is concerned with seed selection. Wrong choice of seed variety cannot be compensated with fertilisation, spraying or the use of mechanisation later in the season. The purpose of this work was to design the strategy for selecting soybean varieties that should be planted on the test farm in order to maximise yield in the following season, based on the knowledge acquired from heterogeneous historical data. We propose weighted histograms regression to predict the yield of different varieties and compare our method to conventional regression algorithms. Based on the predicted yield, we perform portfolio optimisation to come up with the optimal selection of seed varieties that is to be planted. Presented algorithms and results were produced within the Syngenta Crop Challenge.

1. Introduction

The world's growing food demand (Godfray et al., 2010; Ash et al., 2010) challenges seed industry to develop and improve seed varieties, but also challenges farmers to select appropriate seeds among hundreds of varieties available nowadays (Sperling et al., 2014; McGuire and Sperling, 2016). What farmers would certainly need is a portfolio of seed varieties, customised for the environmental conditions at their farm, which would maximise the yield and reduce the insecurity that comes from its variability (Hanson, 2013). Such a targeted solution is important for both traditional (Yengoh, 2012; Louette et al., 2000) and precision agriculture, where the decisions are made locally, on the smallest possible scale (Gassner et al., 2013). In general, there are numerous parameters that influence crop yield. Most prominent are climate and weather conditions, soil type, seed variety and land management, but in the end, it is their complex interaction that determines the yield.

In order to make the decision which seed varieties would be suitable for the given parcel and its environmental parameters, it is necessary to predict their yields. There is an increasing number of scientific researches dealing with yield prediction of various types of crops, fruit and vegetables. Some are based on image processing like in (Pantazi et al., 2016; Liakos et al., 2015), where yield was predicted using NDVI extracted from satellite images and images acquired by a handheld camera. Another approach is to analyse the physical properties of plants, such as height, grain weight and peduncle length (Romero et al., 2013), number of flowers on apple trees (Aggelopoulou et al., 2011) or chlorophyll content measured with SPAD (Saruta et al., 2013). Weather data can also serve as input for yield prediction (Marinković et al., 2009; Brdar et al., 2011; Gonzalez-Sanchez et al., 2014). For example, rainfall in May and a lot of sunshine in June can positively affect the yield of wheat in Serbia, whereas dry spring and extremely hot June can affect it negatively.

The problem with seed selection is that, no matter how successful they may be, none of the aforementioned in-season methods can be applied. It is impossible to know crop vigor, plant height or even weather conditions for the next year. However, the condition of soil does not change dramatically one year after another. It has been shown that content of organic matter, phosphorus, calcium and other compounds in the soil, as well as its pH value, are good indicators of the amount of yield (Drummond et al., 2003). Furthermore, yield can be also estimated based on the ratio of clay, silt and sand, and soil's shallow electrical conductivity (Papageorgiou et al., 2013).

As for the algorithms used for yield prediction, most common ones are artificial neural networks (ANNs) (Pantazi et al., 2016; Drummond et al., 2003; Freitas et al., 2009; Uno et al., 2005; Kaul et al., 2005), multiple regression (Drummond et al., 2003; Kaul et al., 2005) and regression trees (Romero et al., 2013; Marinković et al., 2009). In this work we propose a novel approach to yield prediction – weighted histograms regression (WHR). We approximate the yield probability density function (PDF) at the test farm by forming a histogram of yield, whose entries are weighted according to similarity between test and training farms.

Weighted histograms are not completely new. They have already been used in image processing for motion tracking

* Corresponding author.
E-mail address: oskar.marko@uns.ac.rs (O. Marko).

http://dx.doi.org/10.1016/j.compag.2016.07.009
0168-1699/

(Comaniciu et al., 2003), where an object's feature PDF needs to be calculated. Pixels in the centre of an object are more reliable and thus are attributed with a higher weight. Peripheral pixels are less reliable due to occlusion and interference from the background, and are thus taken with a lower weight. Also, in object recognition, target objects are compared to objects from the database by colour histograms. Since colour is susceptible to changes caused by varying illumination, similar colours are also taken into account – the more similar they are, the more they will contribute to the histogram (Jia et al., 2006).

Yield prediction is just a step towards seed selection. Having known the values of yield predicted for each seed variety, portfolio optimisation theory comes into play. It is a well established theory, originally used for choosing the right portfolio of investments on stock market, which would maximise the return and minimise the risk (Markowitz, 1952). Lately, there have been some examples of its usage in agriculture, as well. It is usually employed in seed variety selection (Nalley et al., 2009; Nalley and Barkley, 2010; Barkley et al., 2010), where predicted yield corresponds to financial return (Freitas et al., 2009), but there are also cases of its use for e.g. irrigation decision-making in condition of reduced water availability (Paydar and Qureshi, 2012), forest planning under the effects of climate change (Dragicevic et al., 2016) and for selecting optimal mix of tree families (Weng et al., 2013). It is always a good strategy to grow plants that respond differently to different environmental conditions and thus statistically better cope with weather unpredictability (Di Falco, 2012). This is especially important for ensuring yield stability in low-income nations and increasing drought and pest tolerance of crops (Barkley et al., 2010).

Whereas high prediction accuracy has been achieved only with classification of the yield into categories, such as low, medium and high (Romero et al., 2013; Saruta et al., 2013; Papageorgiou et al., 2013) and with in-season predictions (Marinković et al., 2009; Brdar et al., 2011; Kaul et al., 2005), we show that it is possible to achieve a high accuracy prediction for one year in advance by using the weighted histograms regression approach. By using this method along with convex optimisation and portfolio optimisation theory it is possible to select a portfolio of seeds, which maximises the yield.

2. Data

Algorithms and related results presented in this paper have been in part developed within the Syngenta Crop Challenge (Syngenta Crop Challenge, 2016), where competitors were provided with necessary historical data about soil, yield and soybean varieties used. The dataset contained 34,212 entries with any of 180 seed varieties planted on one of 120 farms located in the American Midwest (Fig. 1). The varieties were represented with anonymised IDs – v_i, where i took 180 values within the range from 1 to 210.

Season, geographic location, soil properties, common practice and other related parameters were given as features and are listed in Table 1. The sources of data were Syngenta's internal database, ISRIC (World Soil Information) (Hengl et al., 2014), CONUS (Soil Information for Environmental Modeling and Ecosystem Management) (Miller and White, 1998), NASS (United States National Agricultural Statistics Service) (Boryan et al., 2011) and FAO (United Nation's Food and Agriculture Organisation) (FAO, 2016). Some features were contained in datasets of two independent sources. Values from multiple sources were treated as separate features and were all used for prediction.

In the preprocessing phase, we detected that there were multiple entries with the same seed variety planted on the same farm in the same year, but with a different value of yield. We merged them and used only the average yield value accordingly, leaving 32,120 entries. In order to avoid bias and provide reliable yield prediction results for one season, we further split the set into training (seasons 2008–2013) and test dataset (season 2014), with 21,121 and 10,999 entries, respectively. We used only year 2014 as the test dataset to maintain the time causality of our approach. Each year the overall yield gets bigger because of the better mechanisation, pesticides and fertilisers used, as well as other improvements in agricultural production and it was crucial to capture this trendline. In this manner we tried predicting yield in previous years as well. However, the available training dataset reduced dramatically for each preceding year. The number of training samples was insufficient to successfully predict the yield in years before 2014.

3. Methodology and theory

3.1. Prediction using weighted histograms

In order to get the idea about a complexity of the given problem, we used the most straightforward approach by checking the correlation between the yield and individual features, but we did not get any meaningful results. There was no direct link between any of the parameters and the yield. Consequently, we proposed a novel method with the underlying principle that the agricultural system is determinative, i.e. with the same environmental conditions, soil characteristics and seed varieties, different farms give the same yield. In other words, when features of any two farms are compared, the more similar the features are, the more likely it is for the farms to have similar yield. The detailed description of the method follows.

The goal of Syngenta Crop Challenge was to choose up to five soybean varieties that should be planted on the so-called "Evaluation Farm" to maximise the yield. Firstly, we chose a soy variety whose yield we wanted to predict at a test farm. The process was repeated for all available varieties. In the following example, the evaluation farm was denoted as F_E and the variety of interest as v_x. Let us assume that there were five instances of planting v_x in the training dataset. Although this particular variety can be planted on the same farm throughout different years, we can assume without the loss of generality that it was planted on five different farms (F_1 to F_5) (Fig. 2).

Next, we considered the similarity between environmental conditions and other properties at training farms where v_x was planted and related properties at the evaluation farm. They were compared according to individual features – one feature at a time. Let us denote an arbitrary feature according to which the similarity was measured as f_i. The example in Fig. 3 shows values of the given feature at different farms, where the superscript indicates the farm it is related to.

Accordingly, distances of training farms from evaluation farm in the feature's space are shown in descending order in Table 2.

The yield at evaluation farm was more likely to resemble the yield at farms whose value of f_i was closer to f_i^E. Likewise, we could not expect the yield at the evaluation farm to correspond to the yield of a training farm if they had completely different f_is. Another way of explaining this is to view the training farms as advisers, who give their opinion about the yield at the evaluation farm. However, their opinions were not equally important. Opinions of training farms whose f_i was closer to f_i^E were taken with a higher significance than the opinions of those farms whose f_i was far away from f_i^E. Furthermore, a farm's opinion was simply the value of its own yield. It was as if the training farms were telling the evaluation farm that it would have the same yield as them, but the evaluation farm valued their opinions according to how far they were with respect to the given feature. In order to quantify