Computers and Electronics in Agriculture 99 (2013) 93–100

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Predictive models for yield and protein content of brown rice using support vector machine

Keisuke Saruta [a], Yasumaru Hirai [b,*], Kodai Tanaka [b], Eiji Inoue [b], Takashi Okayasu [b], Muneshi Mitsuoka [b]

[a] Yanmar Co., Ltd., Maibara 5218511, Japan
[b] Faculty of Agriculture, Kyushu University, Fukuoka 8128581, Japan

ARTICLE INFO

Article history:
Received 23 February 2013
Received in revised form 1 September 2013
Accepted 3 September 2013

Keywords:
Rice production
Yield
Protein content
Predictive model
Support vector machine
Pattern recognition

ABSTRACT

Rice production in Japan is facing problems of yield and quality instability owing to recent climate changes, aging of farmers, and a decrease in the farmer population. Thus, it is becoming important to develop an improved rice production technology that utilizes collected data about rice production rather than relying on the conventional technology that is based on the experience and knowledge of individual farmers. We developed predictive models for yield and protein content of brown rice that can provide useful knowledge to support farmer's management decision-making, utilizing data sets from 47 paddy fields where rice was produced under various environments and management styles. Support vector machines (SVMs) were applied to build the predictive models based on explanatory variables representing the growth and nutrition conditions after the heading stage and the meteorological environment after the late spikelet initiation stage. The models achieved quantitative accuracy that was within approximately 1 t ha^{-1} in yield for 85.1% of the total data sets and within 0.8% in protein content for 76.6% of the total data sets, respectively. Further, patterns of explanatory variables classified in three classes of yield and protein content, which were visualized by the predictive models, were reasonable in terms of knowledge of crop science. We found that the predictive models using SVMs had the potential to describe a relation between yield or protein content and multiple explanatory variables that reflected diverse rice production in actual fields, and could provide useful knowledge for decision-making of top-dressing and basal fertilization.

1. Introduction

Rice production is influenced by various factors such as the production environment (e.g., water, soil and meteorology), the farmer's management and conditions of rice growth and nutrition. Conventionally, farmers have interpreted the effects of these factors on rice production based on their experience and knowledge accumulated over many years and made management decisions based on these interpretations. However, rice production in Japan is facing problems of yield and quality instability owing to recent climate changes, aging of farmers, and a decrease in the farmer population. Thus, it is becoming important to collect and utilize various data on rice production to achieve high yield and quality, as well as to transfer skills to the next generation of farmers.

Equipment and systems available for data collection for rice production have been studied and developed for various targets, including the meteorological environment (Fukatsu and Hirafuji, 2005), soil environment (Adamchuk et al., 2004), farmer's work history (Liu et al., 2012), rice growth and nutrition (Peng et al., 1995; Stroppiana et al., 2006; Horio and Konya, 2007), and rice yield and quality (Kawamura et al., 2003; Chosa et al., 2006; Hidaka et al., 2011). On the other hand, only a few studies involving data analysis have been conducted to acquire useful knowledge and rules for decision-making to support farmer's management (Roel and Plant, 2004; Roel et al., 2007; Hirai et al., 2012), which means that farmers have not benefited from the various data zcollected by the available equipment and systems. Thus, the production skills of farmers still depend on experience and knowledge of individual farmers, and it results in unstable yield and quality.

Based on this background, our goal was to develop predictive models for rice yield and quality that can provide useful knowledge for farmer's management decision-making. The authors of earlier studies have developed several predictive models regarding rice yield. Horie et al. (1995)developed a predictive model for potential grain yield that uses the product of the harvest index and the dry matter weight under optimal cultivation. In this model, the dry matter weight is determined by the temperature, solar radiation, and CO_2 concentration in the atmosphere. Pirmoradian and Sepaskhah (2006) developed a more practical prediction model for yield by taking the effects of irrigation water and nitrogen application

* Corresponding author. Address: Faculty of Agriculture, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 8128581, Japan. Tel./fax: +81 92 6422956.
E-mail address: hirai@bpes.kyushu-u.ac.jp (Y. Hirai).

0168-1699/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compag.2013.09.003

into consideration, while yield was predicted using the product of the harvest index and the dry matter weight, like the model developed by Horie et al. (1995). In these two models, since the dry matter weight is calculated based on physiological and physical processes of rice growth, the model parameters must be determined via laborious experiments to define the relations between the dry matter weight and meteorological conditions or applications of water and nitrogen. Another drawback is that the model parameters determined under the limited conditions of the experiment don't reflect diverse environments and management styles of rice production in actual fields. Thus, the above models can only predict potential yield or rough trends regarding variation of yield. A predictive model for rice yield is therefore needed that can improve quantitative accuracy and applicability to diverse rice production in actual fields. Further, it is expected that a predictive model should focus on not only yield but also quality, because improvement of rice quality is becoming important in Japan, given the high demands of consumers regarding rice quality and the necessity of enhancing international competitive power.

In this study, we developed predictive models for yield and protein content of brown rice based on pattern recognition using support vector machines (SVMs), assuming the utilization of different data that reflect diverse rice production in actual fields. We selected protein content as an indicator of rice quality, because protein content was closely correlated with palatability of cooked rice (Matsue et al., 2001). First, the predictive models were built using data sets from 47 paddy fields where rice was produced in various environments and under a variety of management styles. The models predicted three classes (low, middle and high) of yield and protein content of brown rice based on explanatory variables representing the growth and nutrition conditions after the heading stage (nitrogen uptake of rice, SPAD readings, and number of spikelets) and the meteorological environment after the late spikelet initiation stage (temperature and solar radiation). The predictive models were evaluated according to their ratio of making correct classification (hereafter, classification accuracy) and their quantitative accuracy. Finally, patterns of explanatory variables classified in three classes of yield and protein content were visualized by the predictive models, and the results were validated based on knowledge of crop science. The patterns of explanatory variables are considered to be useful indicators for farmer's management decision-making because growth and nutrition conditions are regulated through topdressing and basal fertilization taking the meteorological environment into consideration to achieve targets of yield and protein content. The objective of this study was to assess the predictive models using SVMs in terms of their predictive accuracy, applicability to diverse rice production in actual fields, and usefulness in knowledge acquisition for farmer's management decision-making.

2. Materials and methods

2.1. Support vector machines

We thought that several requirements needed to be considered in the development of a predictive model. The first requirement is that a predictive model must be able to express complex relations between multiple explanatory variables and an objective variable (yield or protein content) by using different data collected in actual fields. The second requirement is that the prediction must be robust for variation in input data (e.g., growth and nutrition conditions of rice plants) collected as representative samples in actual fields. To increase prediction robustness, it is reasonable to predict a class with a range in values rather than a specific numeral.

In this study, support vector machines (SVMs) were used for building predictive models for yield and protein content of brown rice, because SVMs were renowned for portraying nonlinear relations between multiple explanatory variables and classes with high generalization performance. The SVMs described in this section are based in Abe (2010) and Platt (1998). An SVM is a binary linear classifier used to separate a set of positive examples from a set of negative examples. SVMs realize high generalization ability by determining the separating hyperplane with the maximum margin. The margin is defined by the distance from the separating hyperplane to the nearest positive and negative examples. An SVM becomes a non-linear classifier by determining the separating hyperplane in a high-dimensional feature space mapped from an input space. When training examples are linearly inseparable, the separating hyperplane is determined by maximizing a soft margin. In soft margin SVMs, constraints in maximizing the margin are relaxed to accommodate linearly inseparable examples. Using techniques of mathematical programming, training of an SVM that separates two classes (G_1 and G_2) is expressed by the following optimization problem on Lagrange multipliers:

$$\begin{aligned}&\max_{\alpha}\left\{\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)\right\}\\ &\text{subject to}\quad \sum_{i=1}^{n}\alpha_i y_i=0,\\ &\qquad 0\leqslant\alpha_i\leqslant C,\\ &\qquad i,j=1,2,\ldots,n\end{aligned}\tag{1}$$

where C is the margin parameter that determines the trade-off between the maximization of the margin and minimization of the classification error, i and j are the indices of training examples, K is a kernel function that maps an input space into a high dimensional feature space and simultaneously calculates the inner product required for training an SVM, n is the number of training examples, $\mathbf{x}_i$ is the input vector, y_i is the output that equals 1 for G_1 or -1 for G_2 and α is the Lagrange multiplier that has a one-to-one relationship with each training example. There are three cases for α_i: (1) $\alpha_i = 0$. In this case, $\mathbf{x}_i$ is correctly classified. (2) $0 < \alpha_i < C$. In this case, $\mathbf{x}_i$ is located near the separating hyperplane and is called a support vector. Thus, the ith training data set contributes to the determination of the separating hyperplane. (3) $\alpha_i = C$. In this case, $\mathbf{x}_i$ is not linearly separated. The classification by the SVM is given as follows:

$$\mathbf{x}\begin{cases}\in G_1 & \left(\sum_{i=1}^{n}y_i\alpha_i K(\mathbf{x}_i,\mathbf{x})-b\geqslant 0\right)\\ \in G_2 & \left(\sum_{i=1}^{n}y_i\alpha_i K(\mathbf{x}_i,\mathbf{x})-b\leqslant 0\right)\end{cases}\tag{2}$$

where b is a bias that is calculated from Lagrange multipliers α.

2.2. Construction of multiclass SVMs by the Decision Directed Acyclic Graph

The Decision Directed Acyclic Graph (DDAG) was used to construct multiclass SVMs, called DAGSVM (Platt et al., 2000). In the DAGSVM, all possible binary SVMs are constructed from training examples, and then classification is executed by list processing based on a decision tree. Fig. 1 shows the DDAG for classification in 3-class SVMs. First, a list with all class numbers as elements is generated for the top decision node ([1, 2, 3]). Then, a test example is evaluated against the decision node that corresponds to the first and last elements of the list (1 vs. 3). When a test example does not belong to one of two classes (e.g., not 3), the class is eliminated from the list, resulting in the new list ([1, 2]). The DDAG proceeds to test the first and last elements of the new list (1 vs. 2). Finally, a test example is classified in one class that remains in the list by