Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-16 (2016) 336–340

Plant Breeding Evaluation with Rank Entropy-based Decision Tree

Xiangyu Zhao *,**, **Zhongqiang Liu** *,***, **Fang Dan** *,****, **Kaiyi Wang** *,**

* *Beijing Research Center for Information Technology in Agriculture, Beijing, China*
** *National Engineering Research Center for Information Technology in Agriculture, Beijing, China*
*** *Key Laboratory of Agri-informatics, Ministry of Agriculture, Beijing, China*
**** *Beijing Engineering Research Center of Agricultural Internet of Things, Beijing, China*

(e-mail: {zhaoxy,liuzq,danf,wangky}@nercita.org.cn)

Abstract: With the rapid development of breeding equipments, large scale breeding becomes possible, massive data evaluation has to be done by software automatically. In this circumstance, this paper tries to evaluate variety with the collected informative data through data mining technologies. According to the characteristics, plant breeding evaluation is considered as an ordinal classification task, and a rank entropy-based decision tree algorithm is proposed to do this classification. The algorithm uses historical trait phenotype and evaluation data to construct decision trees, which are then used to generate evaluations for future cultivars with their trait phenotype. To demonstrate the effectiveness, experiments are carried out on three groups of soybean variety comparison tests (early-maturity, medium-maturity, and green soybean). This work can free breeders from a large number of basic work while increase the efficiency of plant breeding.

Keywords: Breeding Evaluation, Decision Support Systems, Ordinal Classification, Rank Entropy, Crop Phenomics.

1. INTRODUCTION

Variety improvement is important to the agricultural development. The evaluation technologies used to choose superior cultivars from a large number of candidates play a significant role. Conventional evaluations are normally done by breeders using their experience. However, with the rapid development of seed industry, the supply of experienced breeders fails to meet the demand. At the same time, big data and information technologies are developing swiftly. In plant breeding field, several information breeding software has been developed, and mass data of traits and evaluations have been recorded, which offer us opportunities to build connections between trait phenotype and evaluations using data and information technologies. The built connections can be used for future evaluations with similar breeding objects.

There have been a variety of statistical analysis methods used in breeding evaluation, analysis of combining ability (Henderson et al., 1952), crop variety testing (Smith et al., 2015), and other related fields, e.g., selection index theory (Smith, 1936; Sölkner et al., 2008), best linear unbiased prediction (Piepho et al., 2008), technique for order preference by similarity to an ideal solution (Wang et al., 2011), principal component analysis (Kurtanjek et al., 2008; Cao et al., 2014), grey breeding science (Guo and Guan, 2005), and similarity-difference theory (Guo et al., 2010). These methods have made great contributions to the development of breeding evaluation. However, they may unsuit to the stage of large-scale. The key parts of these methods, such as evaluation index and weights, ideal varieties, are always set by professional breeders using their experiences. In the present large scale breeding projects, it becomes quite difficult for the breeders to carry out the analysis in their traditional way, it is even more difficult for a new breeder to acquire the experience and continue the work.

To solve this problem, we introduce the historical plant breeding evaluation results to construct evaluation models, and then use it to generate future evaluating predictions with information technology in this paper. The typical plant breeding evaluation results are {Upgrade (U), retain (R), discard (D)}, which are made by plant breeders based on the trait phenotype of different cultivars. These results can be considered as labels for cultivars while the trait phenotype is considered as attribute. Therefore, the breeding evaluation can be treated as a classification problem. Decision tree induction is an efficient, effective, and understandable technique for classification modeling (Quinlan, 2014, 1986). Therefore, this paper will propose a decision tree algorithm to generate plant breeding evaluation results by utilizing the data of traits and historical

10.1016/j.ifacol.2016.10.062

evaluations. This will free breeders from a large number of basic work while increase the efficiency of breeding.

Conventional decision tree algorithms generate decision rules by splitting measures, such as Gini, chi-square, and Gain ratio (Breiman et al., 1984). They have been widely used in the fields of machine learning and data mining. Unfortunately, these algorithms cannot be used in plant breeding evaluation task directly since there are also ordinal information in the evaluation results as well as classification information. Therefore, this evaluation task is more like an ordinal classification task than a classical one. Furthermore, there exists a constraint that the evaluation result should be a monotone function of trait performance of individual cultivars. Considering two cultivars from an experiment, if one gains better performance than the other at every trait, e.g., yield, thousand grain mass, race-specific resistances in early generations etc., the former always gets better evaluation results. In this circumstance, the plant breeding evaluation problem is much like an ordinal classification with monotonicity constraints.

Recently, there are some literatures focusing on solving ordinal classification with monotonicity constraints. Ben-David proposes an attribute-selection metric which takes both the error as well as monotonicity into account while building decision trees (Ben-David, 1995). Feelders and Pardoel propose a collection of pruning techniques to make a non-monotone tree monotone by additional pruning steps (Feelders and Pardoel, 2003). Van De Kamp et al. proposed an algorithm for learning isotonic classification trees by relabeling non-monotone leaf nodes (Van De Kamp et al., 2009). Xia et al. extended Gini impurity to ranking impurity, and adapt it in CART to ordinal classification (Xia et al., 2008). These algorithms mentioned above enhance the capability of extracting ordinal information, and showed the effectiveness of ordinal classification. However, they cannot ensure a monotone tree is built even if the training data are monotone (Hu et al., 2012).

Kotłowski and Slowinski use a nonparametric classification procedure to monotonize data and then generates a rule ensemble consistent with the monotonized data set (Kotłowski and Słowiński, 2009). Greco et al. generalized the classical rough sets to dominance rough sets for analyzing multicriteria decision making problems by replacing equivalence relations with dominance relations (Greco et al., 1999). These algorithms should generate monotone rules for ordinal classification, and offer significant benefit over nonmonotonic ones. However, numerical experiments showed that the ordinal classifier were statistically indistinguishable from nonordinal counterparts, sometimes performed even worse than nonordinal ones (Ben-David et al., 2009). This unexpected phenomenon is due to the high levels of nonmonotonic noise. In fact, in monotonic classification, decisions are usually given by different decision makers at different time, thus lots of inconsistent samples should exist in data sets. Especially in plant breeding field, evaluation results are always generated by different breeders. This makes that the existence of noises is a normal phenomenon. Therefore, the ordinal classification algorithms cannot be sensitive to noisy samples.

Rank entropy-based decision tree for monotonic classification (REDT), which is insensitive to noises, is an effective and robust ordinal classification algorithm (Hu et al., 2012). It will be used to generate plant breeding evaluations in this paper. REDT extends Shannon's information entropy to rank entropy with dominance sets, then uses rank mutual information (RMI) to evaluate the ordinal consistency between random variables, and finally construct decision trees based on RMI. Since different breeding object always leads to different evaluation strategy, the REDTs for plant breeding evaluation will be constructed separately according to individual breeding object. In addition, even for the same breeding object, the data from various tests must be standardized to make sure the data in one decision tree are comparable [1] . Finally, the constructed decision trees will be used to generate breeding evaluating predictions having similar breeding object.

2. METHOD

2.1 Rank Entropy and Rank Mutual Information

As the key part of REDT, rank entropy is used to generate splitting rules. It will be introduced in this subsection.

Let $C = \{x_1, \ldots, x_n\}$ be a set of cultivars and T be a trait phenotype set to describe the cultivars; E is a finite ordinal set of evaluations. The trait $t \in T$ of x_i is denoted by $v(x_i, t)$, and the evaluation of x_i is denoted by $v(x_i, E)$. The ordinal relations between cultivars in terms of trait t or E is denoted by $\leq$. We say x_j is no worse than x_i in terms of t or E if $v(x_i, t) \leq v(x_j, t)$ or $v(x_i, E) \leq v(x_j, E)$, denoted by $x_i \leq_t x_j$ and $x_i \leq_E x_j$, respectively. Given $T^{'} \subseteq T$, we say $x_i \leq_{T'} x_j$ if for $\forall t \in T^{'}$, we have $v(x_i, t) \leq v(x_j, t)$. Accordingly, we can define the dominance set in ordinal constraint:

$$[x_i]^{\leq}_{T'} = \{x_j \in U \mid x_i \leq_{T'} x_j\},$$
$$[x_i]^{\leq}_{E} = \{x_j \in U \mid x_i \leq_{E} x_j\}.$$

Plant breeding evaluation task is to generate evaluation results for cultivars with their trait data. The evaluation rule is a function

$$f : C \to E,$$

which assigns a decision in E to each cultivar in C. To monotonically ordinal classification tasks, the function should satisfy the following constraint:

$$x_i \leq x_j \Rightarrow f(x_i) \leq f(x_j), \forall x_i, x_j \in C,$$

while the constraint of a classical classification task is:

$$x_i = x_j \Rightarrow f(x_i) = f(x_j), \forall x_i, x_j \in C.$$

However, in real-world applications, there is no such function due to the noise and uncertainty of data sets for either classical or monotonically ordinal classification tasks. To solve this problem, Shannon's information entropy is proposed for classical classification tasks, and has been demonstrated to be a outstanding and robust measure (Shannon, 1948). Rank entropy is a extension of Shannon's information entropy in ordinal constraint. It is defined as:

$$RH^{\leq}_{T'}(C) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\left|[x_i]^{\leq}_{T'}\right|}{n} \tag{1}$$

[1] In this paper, the data are standardized according to check cultivars. All the data introduced in the remainder part are standardized data with no more special description.