

doi:10.6041/j.issn.1000-1298.2015.S0.045

基于序相关的作物育种评价性状特征选择方法^{*}

刘忠强^{1,2} 赵向宇³ 王开义⁴ 李民赞¹

(1. 中国农业大学现代精细农业系统集成研究教育部重点实验室, 北京 100083;
2. 国家农业信息化工程技术研究中心, 北京 100097; 3. 农业部农业信息技术重点实验室, 北京 100097;
4. 北京市农业物联网工程技术研究中心, 北京 100097)

摘要: 将育种家对作物性状表现的综合评价融入作物育种评价中,提出了一种基于序相关的作物育种评价性状特征选择方法。首先,从育种数据中筛选训练样本集及候选性状特征集合,计算候选性状特征集合中各性状与育种材料评价结果的相关性和作物在性状特征间表现的相似性,然后,利用计算的相关系数同时综合考虑性状表现的相似性系数,以期选择的性状特征的相关性最大、相似性最小为目标,建立基于序相关的作物育种评价性状特征选择模型。该模型可为不同育种目标提供重点关注的性状特征集合,为数据化育种提供支持及依据。利用该方法对2013年大豆品系鉴定试验中早熟、中熟、毛豆3类育种材料进行了性状特征选择试验,结果验证了方法的有效性。该方法可以作为育种评价方法的前置步骤,与现有的综合评价、模糊综合评价等方法结合作为性状特征权重的确定手段,提高权重确定的科学性。

关键词: 信息化育种 育种评价 特征选择 序相关

中图分类号: TP319 **文献标识码:** A **文章编号:** 1000-1298(2015)S0-0283-07

Phenotype Feature Selection for Crop Breeding Evaluation Based on Ranking Relevance

Liu Zhongqiang^{1,2} Zhao Xiangyu³ Wang Kaiyi⁴ Li Minzan¹

(1. Key Laboratory of Modern Precision Agriculture System Integration Research, Ministry of Education, China Agricultural University, Beijing 100083, China
2. National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
3. Key Laboratory of Agri-informatics, Ministry of Agriculture, Beijing 100097, China
4. Beijing Engineering Research Center of Agricultural Internet of Things, Beijing 100097, China)

Abstract: Traditional breeding evaluation methods focus on information of crop traits, while ignoring the previous evaluation results. In order to enhance the efficiency of material evaluation under the condition of large-scale breeding, the comprehensive evaluation of crop traits was integrated into the breeding evaluation, and a method of phenotype feature selection for crop breeding evaluation based on ranking relevance was proposed. Firstly, the training sample set and the candidate feature set were selected from breeding data, and the correlation between the phenotype feature and the results of evaluation and the similarity of the agronomic traits were calculated. Then, considering the characteristics of the maximum correlation coefficient and the minimum similarity, a model of phenotype feature selection for breeding evaluation based on ranking relevance was constructed. The established model can be used for different breeding objectives focusing on collection of characters, and validity of the model was verified by using three kinds of soybean identification trial in 2013. The model also can be used as the preprocess of

收稿日期: 2015-10-28 修回日期: 2015-11-15
^{*} 北京市科技计划资助项目(D151100004215002、D151100004215004)
作者简介: 刘忠强, 博士生, 主要从事农业信息技术研究, E-mail: liuzq@nrcita.org.cn
通讯作者: 李民赞, 教授, 博士生导师, 主要从事精细农业技术研究, E-mail: limz@cau.edu.cn

breeding evaluation to determine the weights of the traits accurately.

Key words: Information breeding Breeding evaluation Feature selection Ranking relevance

引言

种业是农业的芯片,是促进我国农业长期稳定发展、保障国家粮食安全的根本。作物育种技术是种业的核心,通过改良作物的遗传特性,选择培育高产优质品种,是种业创新和发展的根本动力。作物育种评价从培育品种中评价、选择满足育种目标的优良品种,是作物育种技术的重要环节,是对育种效果的有效保证。

随着我国粮食生产形势的转变,作物育种评价相关技术也发生了巨大变化,从以产量为主^[1]的评价方式向综合考虑产量、品质、抗逆性等因素^[2-4]转变。与此同时,信息技术的飞速发展带动了数据技术快速提升。在育种领域,杜邦先锋、孟山都、先正达等跨国种业企业均投入重金构建了依托商业化育种软件的育种管理体系^[5],实现全球百万数量级的育种材料试验,为大规模育种数据的高效处理和充分利用提供了保障,极大地提升了品种产出率。在国内,也涌现出以“金种子育种平台”为代表的商业化育种软件,正在探索对育种数据的信息化管理与应用技术。这些新技术将带来大量育种数据的信息化积累,为数据化育种、信息化育种的发展带来契机。在此基础上,充分利用现代信息技术,有效结合已有作物育种技术的综合性评价方法已成为作物育种评价技术的一个重要研究方向。

目前,作物育种评价技术已经从依靠育种家经验为主的阶段向数据化、信息化的育种评价阶段转变。方差分析是一种经典的数据分析方法,可以针对单一性状进行作物评价,应用广泛但综合能力较差。主成分分析^[6-7]、模糊综合评价^[8-10]、灰色多维综合评价^[11-13]、同异分析^[14-16]、DTOPSIS^[4,17-18]等方法可以考虑作物的综合表现,已广泛应用在小麦^[9,14,16]、玉米^[10,17]、大豆^[4]、红花^[13]等多种作物的育种评价中,为推动育种技术的发展做出了突出贡献。以上方法在设定指标权重或者性状理想值时,往往需要专家通过显性描述的方式提供专家经验,普遍存在育种经验定量描述合理性难保证、工作量大等问题,无法满足种业迅速发展背景下的大量材料快速初选任务。本文旨在研究一种隐性利用育种家在育种过程中累积的专家经验,并将其用于后续育种评价中的方法,相较育种家主动描述经验更为合理,便于大规模推广和应用。

1 研究目标

作物的性状表现是育种家选择、评价作物的主要依据。因此,作物的性状表现与评价结果的相关性可以反映出该性状对育种家评价作物的影响力,构建描述这种影响力的有效模型是研究作物育种评价技术的可行方向,这是本文的研究目标。典型的育种评价结果是代表不同评价等级的离散型描述语言,用于描述不同级别间育种材料的优劣程度。例如,{升,留,汰}3级评价结果,分别代表当前育种材料可以进入下一育种阶段,继续进行当前育种阶段,或者被淘汰。这样的评价结果可以转换为育种材料优劣的排序关系,即被评为“升”的资源优于评为“留”的,评为“留”的优于评为“汰”的。与此类似,依据性状表现同样可以对育种材料进行排序。以上2类排序结果的相关程度是性状特征与评价结果的序相关性。本文使用这种序相关性对性状特征与评价结果的关系进行建模,并将其用于作物育种评价中的性状特征选择环节。

此外,由于环境、生理、基因等多重因素的影响,不同性状之间会受到彼此的影响而导致趋同或趋异,即性状特征相似性。性状特征相似性会影响性状特征选择的效果,如果选择出的性状特征组合内部相似性过大,则会在育种评价时放大这类因素的影响,降低评价结果的有效性。本文综合考虑作物的性状表现与评价结果的相关程度以及各性状表现间的相似程度,以选择出的性状组合满足性状组合与评价结果的相关性最大且性状组合内部的相似性最小为目标,研究基于序相关的作物育种评价性状特征选择方法。

2 研究方法

针对具体的育种目标,育种家往往会考虑作物的产量、品质、抗逆性等多种因素,依据其性状的综合表现给出不同的评价结果。本文针对不同育种目标充分利用评价结果,计算评价结果与作物性状表现间的相关程度,并依此进行性状特征的选择,辅助育种家在育种及评价时进行更有针对性的工作。

2.1 数据定义

在作物育种评价的过程中,不同的育种目标会侧重不同的性状特征;不同育种材料中的评价结果间不具有直接可比性。因此,本文在进行数据定义时考虑以上因素,将育种数据定义为包含育种材料

编号 n 、所属育种材料 e 、育种目标集合 G 、性状特征集合 T 以及育种材料评价结果 r 的 5 元组 $\langle n, e, G, T, r \rangle$ 。其中, 育种材料编号 n 是各育种材料在育种材料中的编号值, 用于标识不同育种材料对应的数据; 不同育种材料间存在育种材料环境、育种材料目标等因素的不同, 其结果不具可比性, 所属育种材料 e 是对育种材料评价结果作用域的限定; 育种目标集合 G 是当前育种材料所针对的育种目标, 其值可以为 1 或多个育种目标的集合; 性状特征集合 T 以键值对形式表示当前育种材料对应各性状特征的取值; 育种材料评价结果 r 是当前材料在育种材料中的评价结果, 典型的育种材料评价结果包含 {升, 留, 汰} 3 类, 分别代表当前材料可以进入下一育种阶段, 继续进行当前育种阶段, 或者被淘汰。一条 5 元组数据样例为: $\langle \#1, \text{大豆品鉴}, \{\text{生育期}: 102.0, \text{小区产量}: 0.660, \text{品鉴亩产}: 146.75, \dots, \text{结荚习性}: \text{亚}, \text{花叶病毒盛花期}: \text{免疫}\}, \text{E1}, \text{汰} \rangle$ 。

2.2 数据预处理

在作物育种评价的过程中, 2.1 节中介绍的育种数据的 5 元组表示方式, 其中 n, e, G 为标识型数据, 主要用于对数据进行区分; T, r 是量化型数据, 是本文进行性状特征选择时计算的主要数据, 为保证计算的效率和效果, 需要对其进行数据预处理。

育种材料评价结果 r 包含育种家对育种材料的评价结果, 通常为离散型的描述语言, 代表不同的评价等级, 因此, 可以对其取值进行序数化, 以便计算不同性状与评价结果间的序相关性。例如 {升, 留, 汰} 3 类结果可以序数化的表示为 {3, 2, 1}。

性状特征集合 T 中包含多种类型、多种单位、多种来源的数据, 为了便于计算需进行包括规范化、去噪、归一化的预处理过程, 具体内容分别介绍如下:

(1) 规范化。对性状特征集合 T 中每种性状特征进行规范化处理, 包括统一量化方式、统一计量单位以及统一表现形式。其中, 统一量化方式主要针对枚举型性状, 需依其实际意义使用序数值进行量化表示, 例如大豆的花叶病毒盛花期性状主要包含 {免疫, 高抗, 抗, 感} 4 种性状值, 可使用 {9, 7, 5, 1} 对其进行量化表示。统一计量单位主要针对数值型性状, 如果某性状不同数据的计量单位不一致, 需换算成统一的计量单位, 例如将以 m、cm 等单位计量的株高性状都统一换算为以 m 计量。统一表现形式针对有多表现形式的性状进行统一, 例如日期型性状, 既可以表示为 yyyy-mm-dd, 也可以表示为 mm-dd-yyyy 等其它形式, 将其皆统一使用 yyyy-mm-dd 表示。

(2) 去噪。主要包括参考值域去噪和离群点去噪 2 类。参考值域去噪将不在预设的性状参考值范围内的性状特征所对应的数据剔除, 例如, 大豆的株高性状参考值为 30 ~ 180 cm, 需将株高大于 180 cm 或小于 30 cm 的数据剔除; 离群点去噪通过寻找并剔除大幅偏离数据范围的离群点数据, 降低数据噪声, 离群点数据定义为

$$\{T_i(j) \in T_i \mid |T_i(j) - \bar{T}_i| > K\sigma\} \tag{1}$$

式中 T_i ——性状特征集合 T 中第 i 个性状特征
 $\bar{T}_i$ —— T_i 的平均值
 σ —— T_i 的标准差
 K ——预设参数
 $T_i(j)$ ——满足定义的离群点数据

(3) 归一化。对经过去噪处理之后的性状特征集合 T 进行归一化处理, 归一化方式为

$$T'_i = \frac{T_i - \min T_i}{\max T_i - \min T_i} \tag{2}$$

式中 T'_i —— T_i 经归一化处理后的结果

2.3 性状特征选择步骤

综合考虑作物的性状表现与评价结果的相关程度以及各性状表现间的相似程度, 以选择出的性状组合满足性状组合与评价结果的相关性最大且性状组合内部的相似性最小为目标, 提出作物育种评价性状特征选择方法。该方法的具体步骤如图 1 所示。

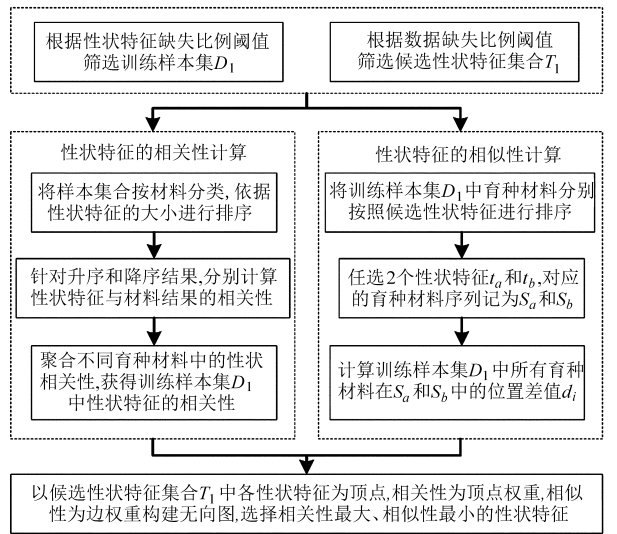


图 1 方法流程图

Fig. 1 Flow chart of proposed approach

2.3.1 筛选训练样本集与候选性状特征集合

本文提出的性状特征选择方法是以育种目标为基础, 不同的育种目标侧重的性状特征差异较大。因此, 在进行特征选择时需首先确定育种目标, 并从育种数据中获取该目标对应的数据集 D_0 。接着, 从 D_0 中筛选训练样本集及候选性状特征集合, 筛选

步骤如下：

(1) 根据数据集 D_0 ，从性状特征集合 T 中获取性状特征并集 T_0 ，并统计 T_0 中性状特征 t_i 的数据缺失比例

$$M_i = \frac{\text{card}(\{d_k \in D_0 \mid d_k[i] = \emptyset, t_i \in T_0\})}{\text{card}(D_0)} \quad (3)$$

式中 d_k —— D_0 中任意数据
 $d_k[i]$ ——数据 d_k 对应性状特征 t_i 的数值
 $\text{card}()$ ——集合中元素的个数

(2) 根据数据缺失比例阈值 M_d 筛选候选性状特征集合 T_1

$$\{t_i \in T_1 \mid M_i \leq M_d\} \quad (4)$$

式中 M_d ——依据专家经验设定的数据缺失比例阈值

当数据缺失比例大于 M_d 时，可认为该数据为无效数据，本文设定 M_d 取值为 20%。

(3) 根据候选性状特征集合 T_1 ，统计数据集合 D_0 中数据 d_k 的性状特征缺失比例

$$M_k^* = \frac{\text{card}(\{t_i \in T_1 \mid t_i[k] = \emptyset, d_k \in D_0\})}{\text{card}(T_1)} \quad (5)$$

式中 $t_i[k]$ ——性状特征 t_i 对应数据 d_k 的数值

(4) 根据性状特征缺失比例阈值 M_T 筛选训练样本集 D_1

$$\{d_k \in D_1 \mid M_k^* \leq M_T\} \quad (6)$$

式中 M_T ——依据专家经验设定的性状缺失比例阈值

当性状缺失比例大于 M_T 时，可认为该性状为无关性状，本文设定 M_T 取值为 20%。

(5) 对训练样本集 D_1 中的数据统计候选性状特征集合 T_1 中性状特征 t_i 的数据缺失比例

$$M'_i = \frac{\text{card}(\{d_k \in D_1 \mid d_k[i] = \emptyset, t_i \in T_1\})}{\text{card}(D_1)} \quad (7)$$

如果 D_1 中所有数据皆满足 $M'_i \leq M_d$ ，则 D_1 为训练样本集合， T_1 为候选性状特征集合；否则，令 $D_0 = D_1$ ， $T_0 = T_1$ ，重复步骤 (2) ~ (5)，继续筛选训练样本集合和候选性状特征集合。

2.3.2 性状特征相关性计算

通过计算候选性状特征集合中各性状与育种材料评价结果的相关性，可以利用育种家在历史育种材料中的专家经验，为性状特征的选择提供依据。根据 2.2 节数据预处理中介绍的方法，使用排序相关性描述性状特征和评价结果的相关程度。此外，不同育种材料间存在育种材料环境、育种材料目标等因素的不同，其结果不具可比性，在计算性状特征相关性时需首先在同一育种材料内部进行，然后对不同育种材料的相关性进行聚合。性状特征相关性

使用 NDCG^[19] 描述，具体计算过程如下：

(1) 根据训练样本集 D_1 ，得到所属育种材料 e 的个数 k ，并将 D_1 中的数据按所属育种材料 e 划分到 k 个子集，每个子集中的数据是同一次育种材料的数据，记为 H_i ， $1 \leq i \leq k$ 。

(2) 从候选性状特征集合 T_1 中选择性状特征 t_j 。

(3) 将 H_i 中的数据依据性状特征 t_j 的大小进行排序，将升序排序结果记为 H_i^0 ，将降序排序结果记为 H_i^1 。

(4) 根据育种材料评价结果 r ，分别对所述 H_i^0 和 H_i^1 中的数据进行微调，以保证相应集合中 t_j 取值相同的数据，其排序关系与相应评价结果的降序关系一致。

(5) 分别针对 H_i^0 与 H_i^1 ，计算性状特征 t_j 与育种材料评价结果 r 的相关性，计算公式为

$$\omega(j, i) = \frac{f(j, i)}{f^*(j, i)} \quad (8)$$

其中
$$f(j, i) = \sum_{p=1}^m \frac{2^{r(p)} - 1}{\text{lb}(1 + p)} \quad (9)$$

$$f^*(j, i) = \sum_{p=1}^m \frac{2^{r^*(p)} - 1}{\text{lb}(1 + p)} \quad (10)$$

式中 m —— H_i 中育种材料的编号
 $r(p)$ —— H_i^0 或 H_i^1 中第 p 位置的育种材料对应的评价结果
 $r^*(p)$ —— H_i 中育种材料依据评价结果降序排序时第 p 位置对应的育种材料评价结果

$\omega(j, i)$ 依其使用 H_i^0 或 H_i^1 中的排序结果分别记为 $\omega^0(j, i)$ 或 $\omega^1(j, i)$ 。

(6) 聚合不同育种材料中的性状相关性，获得训练样本集 D_1 中性状特征 t_j 的相关性

$$\omega(j) = \max\{\omega^0(j), \omega^1(j)\} \quad (11)$$

其中
$$\omega^0(j) = \frac{\sum_{i=1}^k \text{card}(H_i) \omega^0(j, i)}{\sum_{i=1}^k \text{card}(H_i)} \quad (12)$$

$$\omega^1(j) = \frac{\sum_{i=1}^k \text{card}(H_i) \omega^1(j, i)}{\sum_{i=1}^k \text{card}(H_i)} \quad (13)$$

重复步骤 (2) ~ (6) 直至完成候选性状特征集合 T_1 中所有候选性状特征的相关性计算。

2.3.3 性状特征相似性计算

作物不同性状之间存在着相关关系，会导致这些性状有趋同或趋异的表现。例如水稻的理论产量是由每穗实粒数、千粒质量、有效穗数 3 个性状数值

计算获得。如果在进行作物评价时重复考虑以上性状表现,则会降低评价结果的准确程度。因此,计算作物在性状特征间表现的相似性,以此描述性状特征之间的关系。与计算性状特征相关性类似,在计算性状特征相似性时亦使用排序相似性。典型的描述排序相似性的方法包括斯皮尔曼相关系数、肯德尔系数^[20-22]等。本文使用斯皮尔曼相关系数计算,具体计算过程如下:

(1) 将训练样本集 D_1 中育种材料分别按照候选性状特征集合 T_1 中各性状特征进行排序。

(2) 从候选性状特征集合 T_1 中任选 2 个性状特征 t_a 和 t_b , t_a 对应的育种材料序列记为 S_a , t_b 对应的育种材料序列记为 S_b 。

(3) 计算训练样本集 D_1 中所有育种材料在 S_a 和 S_b 中的位置差值

$$d_i = S_a(i) - S_b(i) \tag{14}$$

(4) 计算性状特征 t_a 和 t_b 的相似性为

$$\text{Sim}(a, b) = \left| 1 - \frac{6 \sum_{i=1}^n d_i^2}{u(u^2 - 1)} \right| \tag{15}$$

式中 u ——训练样本集 D_1 中数据的个数

重复步骤(2)~(4)直至完成所有候选性状特征对的相似性计算。

2.3.4 性状特征相似性计算

2.3.2 节和 2.3.3 节中计算所得的性状特征相关性和性状特征相似性是进行性状特征选择的重要依据。在进行性状特征选择时,期望选择的性状特征的相关性最大、相似性最小。该目标可以通过线性组合进行整合,整合后的目标形式为

$$\begin{aligned} \max \quad & \sum_{i \in T_1} \omega(i) x_i - c \sum_{i \in T_1} \sum_{j \in T_1, j \neq i} \text{Sim}(i, j) x_i x_j \tag{16} \\ \text{s. t.} \quad & \sum_{i \in T_1} x_i = n_t \end{aligned}$$

式中 c ——预定义的平衡参数,权衡选择的性状特征相关性最大和相似性最小在最终优化目标中的重要程度

n_t ——选择的性状特征个数

x_i ——第 i 个性状特征是否被选择的指示函数

当 $x_i = 1$ 时,表示第 i 个性状特征被选择,当 $x_i = 0$ 时,表示它未被选择。

在进行性状特征选择时需对式(16)进行优化,这是一个典型的“0-1 规划”问题,目前缺乏快速有效的求解方式,本文使用贪心算法^[23]寻找近似最优解,具体选择过程如下:

(1) 以候选性状特征集合 T_1 中各性状特征为顶点,各性状特征的相关性为顶点权重,顶点之间边

的权重为所述顶点之间的相似性,构建无向图 G_0 。

(2) 构建已选性状特征集合 Y ,使用空集进行初始化,即 $Y_0 = \emptyset$ 。

(3) 从无向图 G_0 中选择权重最大的顶点 v_k 。

(4) 根据 v_k 与其它顶点间边的权重,更新无向图 G_0 中其它顶点权重值,更新方法为

$$\omega_j \leftarrow \omega_j - 2c \text{Sim}(k, j) \quad (j \neq k) \tag{17}$$

(5) 将 v_k 从无向图 G_0 中移至已选性状特征集合 Y ,则有

$$Y_{i+1} = Y_i \cup \{v_k\} \tag{18}$$

$$G_{i+1} = G_i \setminus \{v_k\} \tag{19}$$

判断已选性状特征集合 Y 中是否已有预设个数的性状特征,若有,则将当前集合 Y_i 作为结果输出;若不足,则重复步骤(3)~(5),继续选择新特征。

3 试验分析

3.1 试验设置

使用 2013 年某种业公司大豆品系鉴定试验中早熟、中熟、毛豆 3 类育种材料的真实数据进行分析。这 3 类育种材料同属大豆品种,但其选育的育种目标不同,侧重的性状特征各异。其中,毛豆类育种材料相较其他 2 类资源,更侧重口感等影响其经济效益的选育目标,因此,其关注的性状理应与另外 2 类区别更大。本试验将 3 类育种材料分别视为 3 个不同的育种目标,每个目标各包含一组育种材料。

3.2 结果与分析

对完成预处理后的数据使用 2.3.1 节中介绍的方法进行训练样本集和候选性状特征集合的筛选。筛选过后,3 类材料分别有 121、396、51 组有效数据被选入训练样本集,皆包含生育期、小区产量、品鉴亩产、百粒质量、株高、播种期、出苗期、出苗势、开花期、成熟期、叶形、花色、结荚习性、倒伏性、主茎节数、结荚高度、分枝数、有效荚数、单株粒数、单株粒质量、盛花期花叶病毒共 21 组候选性状特征。

使用 2.3.2 节中介绍的方法计算性状特征的相关性结果如表 1 所示。表中针对不同育种目标,分别对其相关性最大的 5 个性状特征加 * 表示,其中由于存在并列排名第 5 的性状,中熟组有 6 个性状被加 * 表示。

观察表 1 数据,可以发现被加 * 3 次的有小区产量、品鉴亩产、开花期、单株粒质量 4 个性状特征,占总数的 80%。这说明单从性状特征的相关性角度考虑,3 类育种目标具有较强的一致性。但是,如果只考虑排名第 1 的性状特征,早熟组和中熟组更关

表 1 性状特征相关性
Tab.1 Relevance of traits

性状特征	熟期		
	早熟	中熟	毛豆
生育期	0.882	0.823	0.891
小区产量	0.951 *	0.981 *	0.888 *
品鉴亩产	0.951 *	0.981 *	0.888 *
百粒质量	0.844	0.838	0.827
株高	0.871	0.850	0.862
播种期	0.780	0.849	0.749
出苗期	0.780	0.849	0.749
出苗势	0.819	0.886 *	0.843
开花期	0.886 *	0.878 *	0.930 *
成熟期	0.864	0.822	0.838
叶形	0.794	0.730	0.826
花色	0.802	0.870	0.787
结荚习性	0.791	0.668	0.770
倒伏性	0.853	0.863	0.796
主茎节数	0.854	0.837	0.868
结荚高度	0.854	0.812	0.832
分枝数	0.894 *	0.897 *	0.796
有效荚数	0.865	0.865	0.864
单株粒数	0.878	0.874	0.923 *
单株粒质量	0.881 *	0.878 *	0.929 *
盛花期花叶病毒	0.779	0.668	0.750

注小区产量(或品鉴亩产)这类可以直接反映产量的性状特征;毛豆组则更关注开花期,开花期较晚的资源获得的评价较高,另外,这类资源的播种期和出苗期基本一致,这说明毛豆在开花之前经历较长的生长时间可以有效提升其果实的品质。这是直接使用性状特征表现与评价结果的序相关关系进行性状特征选择的结果。

使用 2.3.3 节介绍的性状特征相似性计算方法和 2.3.4 节介绍的性状特征选择方法,取 $n_i = 5$,则性状特征选择的结果如表 2 所示。

表 2 性状特征选择结果
Tab.2 Result of trait selection

序号	早熟	中熟	毛豆
1	小区产量	小区产量	开花期
2	结荚高度	倒伏性	出苗势
3	成熟期	花色	结荚习性
4	倒伏性	出苗势	分枝数
5	出苗势	单株粒质量	结荚高度

观察表 2 数据,可以发现 3 类育种材料共同关

注的性状是出苗势。对比表 1 数据,虽然出苗势仅在中熟组中排入了相关性最大的 5 个性状中,但在 3 类育种材料中相关性都大于 0.8,反映了其表现与评价结果的强相关特性。表 2 中的结果进一步反映了在考虑性状特征之间的相似性之后,出苗势对 3 类资源的重要性。因此,在针对其它育种目标的大豆品种选育过程中,该性状表现可能会具有较大的影响力。此外,在考虑性状特征的相似性进行选择时,表 1 中被共同关注的 4 个性状没有 1 个被 3 类资源共同选择。这说明性状特征的相似性带来了选择结果的差异性。但是,在早熟组和中熟组,小区产量、倒伏性和出苗势被共同选择,占比 60%,而且小区产量是 2 类育种目标最关注的性状。该结果反映出这 2 类育种目标具有较强的一致性。相应地,毛豆组中关注的开花期、结荚习性和分枝数都未出现在早熟组和中熟组的选择结果中,也就是说毛豆组与另 2 类资源的差异性至少为 60%。该结果与预期的 3 类资源的差异性一致。

4 结论

(1) 提出基于序相关的作物育种评价性状特征选择方法,将育种家在作物育种过程中的专家经验进行挖掘应用,针对育种目标,将育种材料的评价结果与作物性状特征的表现建立联系,并兼顾不同性状特征表现间的相似性,进行性状特征的选择。

(2) 对 2013 年某种业公司大豆品系鉴定试验中早熟、中熟、毛豆 3 类育种材料的真实数据进行性状特征选择,发现 3 类资源都关注的性状是出苗势;早熟组和中熟组还共同关注小区产量和倒伏性,其一致性较强;毛豆组作为经济作物,更关注开花期、结荚习性、分枝数 3 类性状。

(3) 基于序相关的作物育种评价性状特征选择方法可以辅助育种家在育种及评价时进行更有针对性的工作,为数据化育种提供支持及依据,是信息化育种技术的有效探索。

(4) 该方法可作为育种评价方法的前置步骤,与现有的综合评价、模糊综合评价等方法结合作为性状特征权重的确定手段,保证权重确定的科学性。此外,还可以在该方法的基础上利用机器学习技术探索基于排序学习的作物育种评价方法,这将是后续研究的重点工作。

参 考 文 献

1 张斯梅,杨四军,顾克军,等. 小麦区域试验产量性状及其稳定性分析[J]. 中国农学通报,2012,28(3):172-176.
Zhang Simei, Yang Sijun, Gu Kejun, et al. Analysis of yield traits and stability in wheat regional trials[J]. Chinese Agricultural Science Bulletin, 2012, 28(3): 172-176. (in Chinese)

- 2 何中虎,夏先春,陈新民,等. 中国小麦育种进展与展望[J]. 作物学报,2011,37(2):202–215.
He Zhonghu, Xia Xianchun, Chen Xinmin, et al. Progress of wheat breeding in China and the future perspective[J]. Acta Agronomica Sinica, 2011, 37(2): 202–215. (in Chinese)
- 3 张灿军,冀天会,杨子光,等. 小麦抗旱性鉴定方法及评价指标研究 I. 鉴定方法及评价指标[J]. 中国农学通报,2007,23(9):226–230.
Zhang Canjun, Ji Tianhui, Yang Ziguang, et al. Study on resistance drought identify method and evaluation index of wheat I identify method and evaluation index[J]. Chinese Agricultural Science Bulletin, 2007, 23(9): 226–230. (in Chinese)
- 4 卢为国,李卫东,梁慧珍,等. DTOPSIS 法综合评价大豆新品种的初步探索[J]. 中国油料作物学报,1998,20(3):22–26.
Lu Weiguo, Li Weidong, Liang Huizhen, et al. A preliminary research on multifactorial evaluation of new soybean varieties with application of DTOPSIS method[J]. Chinese Journal of Oil Crop Sciences, 1998, 20(3): 22–26. (in Chinese)
- 5 董春水,才卓. 现代数字育种技术的研究进展[J]. 玉米科学,2013,21(1):1–8.
Dong Chunshui, Cai Zhuo. Advances in modern data-driven breeding technologies[J]. Journal of Maize Science, 2013, 21(1): 1–8. (in Chinese)
- 6 Cao Y, Jiang Y, Gao H, et al. Development of a model for quality evaluation of litchi fruit[J]. Computers and Electronics in Agriculture, 2014, 106: 49–55.
- 7 贺新,杨培岭,任树梅,等. 基于主成分分析的油葵微咸水调亏灌溉灌水效果评价[J]. 农业机械学报,2014,45(增刊):162–167.
He Xin, Yang Peiling, Ren Shumei, et al. Evaluation of regulated deficit irrigation performance with saline water based on principal component analysis[J]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(Supp.):162–167. (in Chinese)
- 8 曹雯梅,路风银,刘松涛,等. 基于信息熵理论的模糊综合评价方法及其在棉花区试中的应用[J]. 中国农学通报,2013,29(8):212–215.
Cao Wenmei, Lu Fengyin, Liu Songtao, et al. A modified fuzzy synthetic evaluation method based on entropy weight theory and its application in regional test of cotton cultivars[J]. Chinese Agricultural Science Bulletin, 2013, 29(8): 212–215. (in Chinese)
- 9 陈丹花,陆健,孙军勇,等. 应用模糊综合评价法筛选用于啤酒酿造的小麦品种[J]. 食品工业科技,2008,29(8):93–97.
Chen Danhua, Lu Jian, Sun Junyong, et al. Screening wheat variety for beer-brewing by fuzzy evaluation method[J]. Science and Technology of Food Industry, 2008, 29(8): 93–97. (in Chinese)
- 10 刘哲,唐日晶,赵祖亮,等. 黄淮海地区玉米品种适宜性精细区划研究[J]. 农业机械学报,2015,46(8):281–288.
Liu Zhe, Tang Rijing, Zhao Zuliang, et al. Regionalization of maize cultivars cultivated in Huang–Huai–Hai Plain of China [J]. Transactions of the Chinese Society for Agricultural Machinery,2015,46(8):281–288. (in Chinese)
- 11 刘录祥,孙其信,王士芸. 灰色系统理论应用于作物新品种综合评估初探[J]. 中国农业科学,1989,22(3):22–27.
Liu Luxiang, Sun Qixin, Wang Shiyun. A preliminary study on multifactorial evaluation of new crop varieties with the application of the grey system theory[J]. Scientia Agricultura Sinica, 1989, 22(3): 22–27. (in Chinese)
- 12 郭瑞林,关立,张进忠. 作物灰色育种学研究的回顾与展望[J]. 科技导报,2005,23(10):52–55.
Guo Ruilin, Guan Li, Zhang Jinzhong. Review of study on crop grey breeding science and its prospect[J]. Science & Technology Review, 2005, 23(10): 52–55. (in Chinese)
- 13 郭丽芬,张跃,徐宁生,等. 红花种质资源主要数量性状的灰色关联度分析[J]. 西南农业学报,2013,26(1):328–333.
Guo Lifen, Zhang Yue, Xu Ningsheng, et al. Grey relational analysis of main quantity traits of safflower (*Carthamus tinctorius* L.) germplasms[J]. Southwest China Journal of Agricultural Sciences, 2013, 26(1): 328–333. (in Chinese)
- 14 郭瑞林,杨春玲,关立,等. 小麦品种区域试验的同异分析方法研究[J]. 麦类作物学报,2001,21(3):60–63.
Guo Ruilin, Yang Chunling, Guan Li, et al. Study on identical and different analysis method of wheat variety regional test[J]. Acta Tritical Crops, 2001, 21(3): 60–63. (in Chinese)
- 15 王阔,郭瑞林. 同异分析方法在绿豆品种区域试验中的应用研究[J]. 杂粮作物,2004,24(1):15–18.
Wang Kuo, Guo Ruilin. Application of identical and different analysis in mung bean variety regional test[J]. Rain Fed Crops, 2004, 24(1): 15–18. (in Chinese)
- 16 郭瑞林,刘亚飞,王景顺,等. 同异理论及其在小麦育种中的应用[J]. 麦类作物学报,2010,30(5):970–975.
Guo Ruilin, Liu Yafei, Wang Jingshun, et al. Similarity-difference theory and its applications in wheat breeding[J]. Acta Tritical Crops, 2010, 30(5): 970–975. (in Chinese)
- 17 卫勇强,赵保献,雷晓兵,等. 灰色关联度分析和 DTOPSIS 法综合评价玉米新品种[J]. 江西农业学报,2009,21(6):11–14.
Wei Yongqiang, Zhao Baoxian, Lei Xiaobing, et al. Comprehensive evaluation of new corn varieties by gray relational analysis and DTOPSIS method[J]. Acta Agriculturae Jiangxi, 2009, 21(6): 11–14. (in Chinese)
- 18 马雄风,喻春明,汤清明,等. DTOPSIS 法在苧麻品种多因素综合评价中的应用[J]. 中国麻业,2005,27(6):283–285.
Ma Xiongfeng, Yu Chunming, Tang Qingming, et al. Comprehensive judgement of multiple factors on ramie variety with DTOPSIS [J]. China's Fiber and Products, 2005, 27(6): 283–285. (in Chinese)
- 19 Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques[J]. ACM Transactions on Information Systems, 2002, 20(4): 422–446.
- 20 Kendall M G. Rank correlation methods[M]. London: Griffin, 1970.
- 21 Liebetrau A M. Measures of association quantitative applications in the social sciences [M]. Thousand Oaks, CA: Sage Publications Inc., 1983.
- 22 Abdi H. Kendall rank correlation[M]. Salkind N J. Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage, 2007: 508–510.
- 23 Geng Xiubo, Liu Tieyan, Qin Tao, et al. Feature selection for ranking[C]//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07), 2007: 407–414.