IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 20, 2020, accepted August 15, 2020, date of publication August 20, 2020, date of current version September 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018198

Plant Breeding Evaluation Based on Coupled Feature Representation

XIANGYU ZHAO[1,2], YANYUN HAN[1,3], ZHONGQIANG LIU[1,3], SHOUHUI PAN[1,4], AND KAIYI WANG[1,2]

[1]Beijing Research Center for Information Technology in Agriculture, Beijing 100097, 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

Corresponding authors: Xiangyu Zhao (zhaoxy@nercita.org.cn) and Kaiyi Wang (wangky@nercita.org.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2019YFD1101102.

ABSTRACT With the rapid development of improved breeding equipment and information technology, computer-aided decision-making in plant breeding evaluation can help solve the problems associated with high-throughput demand and insufficient experience of breeders in modern large-scale field breeding experiments. Many linear models have made great contributions to the development of breeding evaluation although they are based on a wrong assumption of attribute independence. This paper proposes a unified coupled representation that integrates intra-coupled and inter-coupled relationships to capture the interdependence among quantitative traits by addressing coupling context and coupling weights. Moreover, a hybrid scheme of the linear correlation and ordinal relation is introduced to express the coupling relationship with a preset parameter that balances the contributions so as to capture both relative and absolute performance in cultivar selection and breeding evaluation. A framework that includes data preprocessing, coupled data representation, feature selection, prediction model construction, and assisted decision-making is our overall solution for the plant breeding evaluation task. Experiments on real plant breeding data sets demonstrated the effectiveness of coupled representation for elucidating the quantitative phenotypic traits and the advantages of the proposed plant breeding evaluation algorithm compared with benchmark algorithms.

INDEX TERMS Breeding evaluation, coupled representation, quantitative phenotypic traits, feature selection, decision support systems.

I. INTRODUCTION

Food security is a key issue worldwide because feeding the several billion people on this planet is a serious challenge, especially with the pressure of global environmental change. It has been estimated that crop production must double by 2050 to meet the predicted production demands of the global population [1]. Moreover, land degradation and water contamination, climate change, sociocultural developments (e.g., dietary preference of meat protein), governmental policies, and market fluctuations add uncertainties to food security [2]. Breeding and promotion of quality varieties with high and stable yields is the most important and effective way to guarantee agricultural production and food security, and this is the fundamental driving force underlying seed industry innovation and development [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Justin Zhang.

Breeding and evaluation of plant varieties in the modern seed industry is based on a large amount of information from previously conducted tests and professional knowledge during the multiple life cycles of crops, and this process involves continuous selection from tens of thousands or even hundreds of thousands of individuals. Crop breeding strategies have gradually shifted from yield-based approaches to comprehensive consideration of yield, quality, water conservation, and stress resistance (i.e., drought tolerance, salinity tolerance, disease and pest resistance), to adapt to changes in the environmental pressure, food production situation, and consumption preferences [4]. This shift has brought new requirements for breeding technologies and great progress has been made.

Molecular breeding is an emerging breeding technology that has great potential. The most commonly applied molecular tools for crop breeding are molecular markers, which are variations at the DNA level that are used to track and

monitor specific regions of the genomes during crossing and selection. This approach can provide important information for parental selection, genetic diversity estimation, reducing linkage drags, and genomics-assisted breeding [4]. The applications of marker-assisted selection for crop breeding, including marker-assisted backcrossing, marker-assisted gene pyramiding, marker-assisted recurrent selection, and genome-wide or genomic selection, have increased because of their low cost, high read accuracy, and competing sequencing systems with abundant successes in developing cultivars in rice, maize, legume, horticultural crops, and several long juvenile species [5]. The production and evaluation of genetically modified crops are other active areas of molecular breeding. Because genetic modification (GM) can generate novel variations beyond those that occur naturally, it is considered to be a major contributor for future crop variety innovation [4]. GM is a powerful tool to create novel alleles, promote superior alleles, and remove deleterious effect alleles [6]. However, access to GM technology is currently restricted in many countries owing to political and bioethical issues [4].

Besides genotype data, phenotype data also are an important part of molecular breeding. For example, phenotype data can be used to train prediction models in genomic selection; a single phenotypic cycle can be used to identify markers that are then used in marker-assisted recurrent selection through generations; and phenotypes can be used to verify the effect of transgenic studies [7]. Therefore, detecting associations between genotypes and phenotypes should be a significant component of future crop-breeding programs. However, compared with the large number of genotype studies, studies of field phenotypes lag dramatically, which has limited the ability to dissect genetic variations of crops [6]. Fortunately, the rapid development of information technology and its applications in plant breeding have facilitated the collection of a lot of breeding-related data using information systems. High-throughput phenotype platforms offer efficient, non-invasive, flexible, and low-cost solutions to collect field phenotype data to bridge this gap using technologies such as novel sensors, image analysis and modeling, robotics, and remote control [7]–[11]. In addition, some breeding data management software programs, such as *GoldenSeed*, *NBS*, *AGROBASE*, and *PRISM*, have been developed to manage various types of breeding data, including genotype, phenotype, and environmental data, and breeding information [12]. This makes possible the comprehensive use of multidimensional breeding data. Therefore, breeder-friendly analytical and decision support tools will be very helpful.

Without affordable and easy to use tools, these new molecular breeding technologies are inaccessible for poor-resource breeding programs, and the logistics of sending plant tissue samples for analysis in a timely manner can be prohibitive, especially in under-resourced countries [4]. In such countries, conventional breeding is still the main way to improve crop cultivars. Initially, plant breeders developed cultivars that they perceived to have traits that met their breeding objectives, such as high yield, superior quality, and disease resistance, by observing the phenotypes and selecting cultivars based on intuition and/or experience; multiple trait selection is common in this circumstance. Breeders of different crop species have different strategies for selecting the primary traits of interest, but the main goal is to balance the relative importance of different traits [13]. For example, a new cultivar may have superior yield to other cultivars used in production, but if the higher yielding cultivar has deficiencies associated with disease resistance, or poor heat and drought tolerance, the stability of superior yield cannot be guaranteed and may cause serious production issues under extreme climate conditions.

Therefore, quantitative genetic studies, which have emphasized that information on the inheritance of quantitative traits can be used to plan breeding strategies for cultivar development, were conducted to balance trait selection, infer the types of genetic effects that are important, and determine how selection methods could be modified to enhance cultivar development and enhance germplasm pools [13]. Additionally, a variety of statistical analyses have been introduced, such as variance analysis, selection index theory, best linear unbiased prediction, principal component analysis, association analysis, analytic hierarchy process, grey breeding science, and similarity-difference theory [14], [15].

These methods have made great contributions to the development of breeding evaluation by effectively improving the degree of dataization and informationization of plant breeding evaluation technology through the analysis and utilization of quantitative trait data [16]. Linear models are the most commonly applied statistical approaches to analyze phenotype data based on the assumption of attribute independence [7]. However, this assumption may not be satisfied in breeding programs, in which the traits generally interact and are coupled via explicit or implicit relationships. For example, the theoretical yield of rice is calculated from three traits of real grains per panicle, 1000-grain weight, and effective panicles, which indicates the coupling relationship between the theoretical yield and the other three traits from a theoretical perspective. Table 1 uses the Pearson correlations to describe the relations between six traits from a very early spring maize trial. It can be found that the correlations between different traits varies greatly. Standard yield and plot yield are strongly related (0.88), whereas moisture content is inversely related with standard yield (-0.39), while stalk lodging and root lodging are almost irrelevant (-0.01). This verifies the existence of the coupling relationship among traits from a practical perspective. As a result, the effective expression of the coupling relationship among traits should be the basis of breeding data analysis and decision-making.

Coupled feature representation has been demonstrated to be an effective solution that describes the relationships among quantitative data in many real-world data sets [17], [18]. The coupling relationships include both intra-coupled interactions within an attribute and inter-coupled interactions among different attributes. Both relationships are introduced into the plant breeding evaluation task by decoupling the

TABLE 1. Correlations between six maize traits.

	Stalk lodging	Moisture content	Standard yield	Plot yield	Root lodging	Empty ratio
Stalk lodging	1.00	0.05	-0.07	-0.12	-0.01	0.25
Moisture content	0.05	1.00	-0.39	-0.03	-0.14	-0.06
Standard yield	-0.07	-0.39	1.00	0.88	0.16	-0.17
Plot yield	-0.12	-0.03	0.88	1.00	0.16	-0.23
Root lodging	-0.01	-0.14	0.16	0.16	1.00	0.16
Empty ratio	0.25	-0.06	-0.17	-0.23	0.16	1.00

relationship among traits to solve the independence problem. Conventionally, the intra-coupled and inter-coupled relationships are calculated separately, and then merged while representation. This paper proposes a unified coupled representation that integrates both intra-coupled and inter-coupled interactions among quantitative traits in one form. It provides a concise expression and improves the calculation efficiency. In addition, the linear correlation (e.g., Pearson correlation coefficient) is always treated as suitable for describing the relationships among quantitative data [17]. However, in plant breeding, the ordinal relationship is also important, because plant breeders develop cultivars by selecting individuals from a population that perform better rather than the ones that achieve a specific performance (e.g., an 18 t/hm^2 rice yield). Therefore, a hybrid relationship of the linear correlation and ordinal relation is proposed to express the coupling relationship with a preset parameter that balances the contributions of these two parts. On the basis of this coupled representation, a plant breeding evaluation framework that includes data preprocessing, coupled data representation, feature selection, prediction model construction, and assisted decision-making is proposed as an overall solution for the plant breeding evaluation task.

The following parts of this paper are organized as follows: section II (''Coupled Feature Representation'') introduces the coupled feature representation for plant breeding evaluation; section III (''Plant Breeding Evaluation'') proposes a plant breeding evaluation framework and algorithm; section IV (''Experiments'') demonstrates the effectiveness of the coupled representation and plant breeding evaluation algorithm based on experiment results on real plant breeding data sets and discussion. Finally, section V (''Conclusions'') summarizes the contributions of this paper.

II. COUPLED FEATURE REPRESENTATION

In real-world data, coupling refers to any relationship between two or more aspects, such as co-occurrence, neighborhood, dependency, linkage, correlation, or causality. There are many kinds of coupling layers (including entity coupling, property coupling, context coupling, interaction coupling, and learning coupling) and coupling forms (including serial coupling, causal coupling, synchronous coupling, exclusive coupling, and dependent coupling) [18]. Based on coupled attribute analysis, the coupling relationships of attributes are typically intra-coupled interactions within an attribute and inter-coupled interactions between attributes [19], [20]. In this paper, the coupling relationships of quantitative traits are represented in a similar way.

Because the phenotypic traits of cultivars are affected by both genetic and environmental factors [13], the quantitative data from different trials cannot be directly compared based on environment. The plant breeding evaluation data is split by trials. The data in one trial can be represented as $< C, T, E >$, where $C = \{c_1, \ldots, c_m\}$ is a set of cultivars (m is the number of cultivars); $T = \{T_1, \ldots, T_n\}$ is the quantitative trait set (n is the number of traits) that describes the cultivars, in which $T_j = \{t_1^j, \ldots, t_m^j\}$ includes the quantitative values of all cultivars in C according to trait j, where t_i^j is the value of cultivar c_i according to trait j; and $E = \{e_1, \ldots, e_m\}$ is the evaluation set where e_i is the evaluation result of c_i.

The Pearson correlation coefficient is commonly used to describe the relationship between continuous attributes. For plant breeding evaluation, the Pearson correlation coefficient between two traits can be defined as:

$$Cor(T_j, T_k) = \frac{\sum_{i=1}^{m}(t_i^j - \overline{T_j})(t_i^k - \overline{T_k})}{\sum_{i=1}^{m}(t_i^j - \overline{T_j})^2 \sum_{i=1}^{m}(t_i^k - \overline{T_k})^2}, \quad (1)$$

where $\overline{T_j}$, $\overline{T_k}$ are the respective mean values of T_j, T_k.

However, the Pearson correlation coefficient only evaluates the linear relationship between two traits. To further describe the nonlinear relationship functions, inspired by the idea of Taylor expansion, a Taylor-like series to quantify the global dependency was proposed in [17]. Because any analytic function can be approximated by its Taylor polynomials, the global relationship (both linear and nonlinear) between traits can be represented by the linear correlations between attributes and their extended powers. In this circumstance, the original trait data should be transformed to a Taylor-like form T^L, which can be defined as:

$$T^L = \{T_{11}, T_{12}, \ldots, T_{1L}, T_{21}, T_{22}, \ldots, T_{2L}, \ldots, T_{nL}\}, \quad (2)$$

where L is the maximum expansion power, and T_{jp} indicates the $p-th$ power of the corresponding value of trait j, which can be calculated as:

$$T_{jp} = \{\langle t_1^j \rangle^p, \langle t_2^j \rangle^p, \ldots, \langle t_m^j \rangle^p\}, p \in (1, L), \quad (3)$$

where $\langle t_i^j \rangle^p$ is the $p-th$ power of t_i^j.

Based on the extended trait data T^L, for target trait j, the intra-coupled interaction is quantified as the coupling

relationships between attribute T_j and its powers T_{jp} [17]. This interaction can be defined as:

$$R^a(T_j) = \begin{pmatrix} \theta_{11}(j) & \theta_{12}(j) & \cdots & \theta_{1L}(j) \\ \theta_{21}(j) & \theta_{22}(j) & \cdots & \theta_{2L}(j) \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{L1}(j) & \theta_{L2}(j) & \cdots & \theta_{LL}(j) \end{pmatrix}, \tag{4}$$

where $\theta_{pq}(j) = C(T_{jp}, T_{jq})$ is the coupling relationship between T_{jp} and T_{jq}.

Moreover, the inter-coupled interaction captured the coupling relationships between each trait T_j, which includes $\{T_{j1}, \cdots, T_{jL}\}$, and all the powers of other traits $\{T_k\}_{k\neq j}$ [17]. This interaction can be defined as:

$$R^e(T_j|\{T_k\}_{k\neq j}) = \begin{pmatrix} \eta_{11}(j|k_1)\cdots\eta_{1L}(j|k_1)\cdots\eta_{11}(j|k_{n-1})\cdots\eta_{1L}(j|k_{n-1}) \\ \eta_{21}(j|k_1)\cdots\eta_{2L}(j|k_1)\cdots\eta_{21}(j|k_{n-1})\cdots\eta_{2L}(j|k_{n-1}) \\ \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ \eta_{L1}(j|k_1)\cdots\eta_{LL}(j|k_1)\cdots\eta_{L1}(j|k_{n-1})\cdots\eta_{LL}(j|k_{n-1}) \end{pmatrix}, \tag{5}$$

where $\eta_{pq}(j|k_i) = C(T_{jp}, T_{k_iq})$ is the coupling relationship between T_{jp} and T_{k_iq}.

These intra-coupled and inter-coupled interactions of quantitative traits can be integrated into a unified form, R^u, which can be defined as:

In this unified form, the intra-coupled interaction of trait j is the slice of R^u from $R^u(jL-L+1, jL-L+1)$ to $R^u(jL, jL)$, and the inter-coupled interaction between trait j and trait k is the slice from $R^u(jL-L+1, kL-L+1)$ to $R^u(jL, kL)$.

Based on this unified coupling relationship matrix, the extended trait data T^L can be transformed into the coupled representation data T^c, which can be defined as:

$$T^c = T^L \odot w \otimes R^u, \tag{7}$$

where $w = [w_0, w_0, \cdots, w_0]$ is a $1 \times nL$ vector concatenated by n constant vectors $w_0 = [1/(1!), 1/(2!), \cdots, 1/(L!)]$, which is a $1 \times L$ constant vectors. "$\odot$" denotes the Hadamard product, and "$\otimes$" represents the matrix multiplication.

In general, the coupling relationship between numerical attributes is expressed by the Pearson correlation coefficient, that is, $C(T_j, T_k) = Cor(T_j, T_k)$. However, there are some specific modifications that should be incorporated for plant breeding purposes. Breeders always evaluate the candidate cultivars based on phenotypic trait performance. For example, if there are two cultivars with similar performance but different plot yields (i.e., 5.93 and 4.22 kg), the former cultivar is always considered better because of the bigger plot yield, but this does not consider the correlation between these plot yields. In some circumstances, the ordinal relationship is more important than the correlation for plant breeding evaluation. Therefore, the Kendall rank correlation coefficient, which is a formula for inferring ordinal relationship, is introduced to measure the ordinal association between traits, and can be defined as:

$$\tau(T_{jp}, T_{kq}) = \frac{4P}{m(m-1)} - 1, \tag{8}$$

where m is the number of cultivars and P is the sum, over all the cultivars, of cultivars ranked after the given cultivar based on the values of both traits.

Furthermore, by integrating the linear correlation and ordinal relationship, the coupling relationship can be defined as:

$$C(T_{jp}, T_{kq}) = \alpha \cdot Cor(T_{jp}, T_{kq}) + (1-\alpha) \cdot \tau(T_{jp}, T_{kq}), \quad p, q \in (1, L), j, k \in (1, n), \tag{9}$$

where α is a preset parameter that balances the contributions of the Pearson correlation coefficient and Kendall rank correlation coefficient. The general coupling relationship with the Pearson correlation coefficient is a special case, in which $\alpha = 1$.

So far, the global coupled representation for quantitative traits T^c has been obtained. The findings reflected the mutual influence and interactions of phenotypic traits, and reserved far more coupling relationships than the original representation. It was previously demonstrated that coupled representation may improve classification or cluster performance in many fields [17], and this is further verified for plant breeding in the following sections.

III. PLANT BREEDING EVALUATION

A. EVALUATION FRAMEWORK

Plant breeders observe phenotypes and select cultivars that meet their breeding objectives. Typical plant breeding evaluation results are *upgrade*, *retain*, *discard* [14]. The best cultivars always *upgrade* to the next level, whereas the worst ones are *discarded*, and the common cultivars may be *retained* for further testing.

The evaluation results are made by plant breeders based on the comprehensive performance of phenotypic traits of different cultivars. Therefore, plant breeding evaluation can be considered a classification task, and a plant breeding evaluation framework is proposed (Fig. 1).

Because different traits are measured on different scales and contain some very large outliers, directly utilizing these data may slow down or prevent the convergence of many machine learning algorithms, and even degrade the predictive performance. Therefore, data preprocessing, as the **first step** of the plant breeding evaluation framework, removes outliers and transforms data into a same scale, and is used to ensure the robustness and predictive performance.

The assumption that the individual traits of the crop are independent and identically distributed is not always true. In fact, there is a complex coupling relationship between crop performances. Consequently, as introduced in section II, coupled data representation including data extension and coupling calculation, is used as the **second step** of the framework. After this step, the original quantitative traits are transformed to an nL dimension coupled form that may reflect the mutual influence and interactions of phenotypic traits.

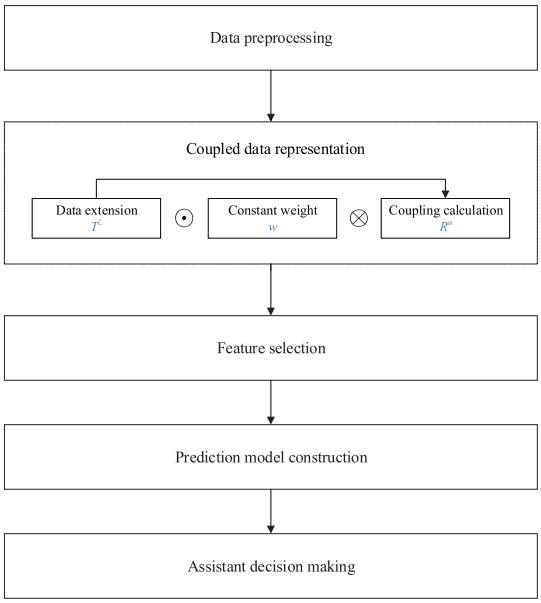


FIGURE 1. Framework of plant breeding evaluation based on coupled feature representation.

However, there might be some redundant or irrelevant information in the coupled representation, which may lead to an over-fitting problem. Therefore, the **third step** is feature selection to capture the key characteristics of the coupled traits for plant breeding evaluation.

The **fourth step** is to train a prediction model by utilizing the selected coupled traits. As plant breeding evaluation is considered a classification task, many classification models can be used to train a prediction model. In this paper, Support Vector Machine (SVM) was used to accomplish this task. It should be noted that, although SVM can handle nonlinear decision boundaries of arbitrary complexity, the linear SVM is used because nonlinear relationships were obtained by the Taylor-like expansion and coupled representation.

Based on the trained model, in the **final step**, a computer-aided decision on unevaluated cultivars can be made with the coupled representation of phenotype data.

B. DATA PREPROCESSING

There are three parts in the data preprocessing step.

The first part involves unifying the quantitative method and the unit of measurement. In plant breeding, some breeders evaluate traits using different quantitative methods. For example, some breeders use {1, 2, 3} to measure the degree of heat tolerance, whereas others prefer {1, 3, 5, 7, 9}. In addition, some traits require a unified unit of measurement; for example, we had to choose *kg* or *g* as the unified plot field unit. This step has been conducted in our established "Golden seed breeding cloud platform" using the trait management module [12].

The second part involves replacing the outliers. Based on practical experience, breeders may predefine the value range of some traits; for example, the plant height of maize is between 50 and 300 *cm*. Thus, values out of this range will be replaced by the corresponding boundary (i.e., t^j_{bmin} for small outliers or t^j_{bmax} for large outliers). Moreover, statistical outliers should also be replaced; that is, for trait j, data meeting the condition of Equ. 10 should be replaced by the corresponding maximum or minimum value defined in Equ. 11:

$$|t^j_i - \overline{T_j}| > K \cdot \sigma_j, \tag{10}$$

$$\begin{aligned} t^j_{min} &= \max(t^j_{bmin}, \overline{T_j} - K \cdot \sigma_j) \\ t^j_{max} &= \min(t^j_{bmax}, \overline{T_j} + K \cdot \sigma_j), \end{aligned} \tag{11}$$

where σ_j is the standard deviation of trait j and K is the specified number of standard deviations to detect outliers.

The third part is normalization, which may scale the data set such that all trait values are in the range [0, 1]. This is a crucial factor in guaranteeing that the plant breeding evaluation model is effectively trained, and is defined as:

$$T'_j = \frac{T_j - \min(T_j)}{\max(T_j) - \min(T_j)}, \tag{12}$$

where $\max(T_j)$ and $\min(T_j)$ are the maximum and minimum values of trait j, respectively.

After these three parts of data preprocessing, the original data are ready for coupled representation (as shown in section II), which is the input for feature selection that will be introduced in the next section.

C. FEATURE SELECTION

The coupled representation of quantitative traits can be used to train the plant breeding evaluation model. However, the dimension of T^c is expected to have a lot of columns because of the multiplication of n and L, and some of the

$$R^u = \begin{pmatrix} C(T_{11}, T_{11}) & \cdots & C(T_{11}, T_{1L}) & \cdots & C(T_{11}, T_{n1}) & \cdots & C(T_{11}, T_{nL}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C(T_{1L}, T_{11}) & \cdots & C(T_{1L}, T_{1L}) & \cdots & C(T_{1L}, T_{n1}) & \cdots & C(T_{1L}, T_{nL}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C(T_{n1}, T_{11}) & \cdots & C(T_{n1}, T_{1L}) & \cdots & C(T_{n1}, T_{n1}) & \cdots & C(T_{n1}, T_{nL}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ C(T_{nL}, T_{11}) & \cdots & C(T_{nL}, T_{1L}) & \cdots & C(T_{nL}, T_{n1}) & \cdots & C(T_{nL}, T_{nL}) \end{pmatrix}, \tag{6}$$

coupled traits have no relationship with the breeding decisions. Therefore, we try to select a smaller number of features to reduce the dimension and capture the characteristics and structure of the coupled representation data to further improve the efficiency of plant breeding evaluation.

Lots of feature selection technologies use feature ranking to select the best features; however, a good feature ranking criterion is not necessarily a good feature subset ranking criterion. Therefore, we use a recursive feature elimination (RFE) [21] to select features by producing a feature subset ranking as opposed to a feature ranking. Feature subsets are nested as:

$$T^s(1) \subset T^s(2) \subset \cdots T^s(n \cdot L), \tag{13}$$

where $T^s(i)$ is a selected subset of coupled traits that contains i trait elements, and $T^s(n{\cdot}L) = T^c$ is the entire set of T^c. After the feature subset selection, the subset that gains the best predictive performance will be selected as the final feature selection result T^s, which can be defined as:

$$T^s = \arg\max_{i \in [1, n \cdot L]} f(T^s(i)), \tag{14}$$

where $f(x)$ is the predictive performance of the estimator, with x as its data set.

Algorithm 1 RFE of Plant Breeding Evaluation

Input:

$< C, T^c, E >$: the coupled traits and evaluations of the cultivars in set C.

Output:

T^s: the selected subset of traits.

1: **Step 1.** initialize the feature subset $T^s = T^s(i) = T^c$ containing $i = n \cdot L$ trait elements, and the global predictive performance $P = 0$;
2: **Step 2.** train the classifier with $T^s(i)$ and E, get the coefficients of traits $c(i)$, and compute the predictive performance $P(i)$;
3: **Step 3.** if $P(i) > P$, update $P = P(i)$ and $T^s = T^s(i)$;
4: **Step 4.** if $i > 1$, get $T_j = \arg\min_{T_j \in T^s(i)} c(i)$, and set $T^s(i-1) = T^s(i) - T_j$;
5: **Step 5.** set $i = i - 1$, and recursively produce Step 2 to Step 5 until $i = 1$;
6: **Step 6.** use T^s as the selected subset of traits.

As shown in Algorithm 1, RFE is used to select traits by recursively considering increasingly smaller sets with an external estimator that computes predictive performance and assigns coefficients of different traits. RFE eliminates some of the coupled traits and retains a minimum subset of features that yields the best classification performance. In this paper, SVM was selected as the estimator and used as the classifier in the fourth step.

IV. EXPERIMENTS

In this section, experiments were performed on several plant breeding data sets; specifically, there were four qualification trials at the T2 level from a breeding company in 2019 (Table 2) to demonstrate the effectiveness of our proposed coupled feature representation and plant breeding evaluation approach.

TABLE 2. Descriptions of data sets.

Data Set	Cultivar	Trait	Short Form
Very early spring maize	135	11	D1
Early spring maize	272	14	D2
Medium spring maize	570	12	D3
Middle late maturing maize	217	13	D4

The experiments included two parts, coupled representation analysis and plant breeding evaluation, to analyze the influence of different parameters on coupled representation and the performance of proposed algorithms compared with several benchmark algorithms.

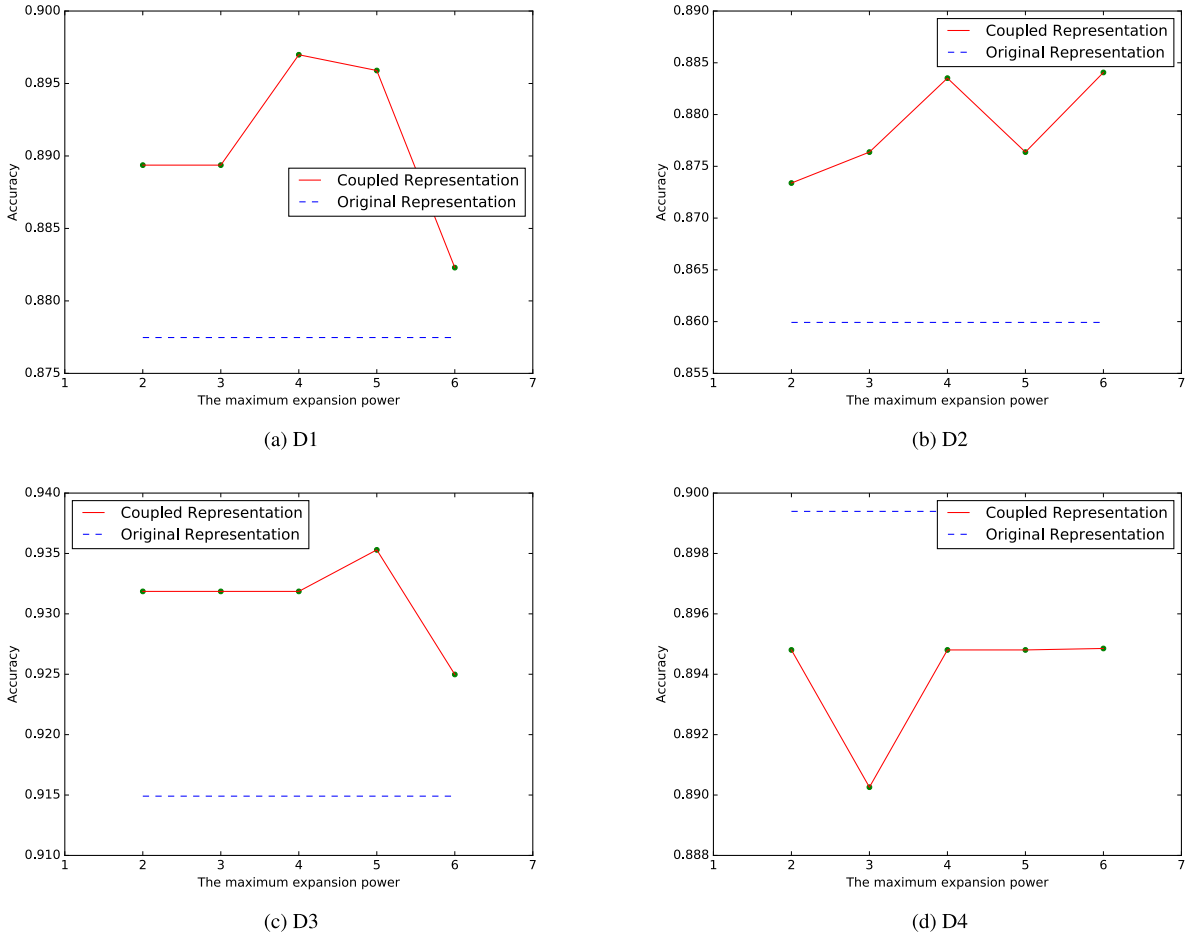
A. COUPLED REPRESENTATION ANALYSIS

Coupled trait representation is key to accomplishing plant breeding evaluation because it reveals the mutual influence and interactions of phenotypic traits. As mentioned in section II, our proposed coupling relationship integrates the linear correlation and ordinal relationship to fit the field specificity in the plant breeding area. In the following experiments, effective approaches for using these two kinds of relationships will be discussed.

As indicated in Equ. 2, Equ. 6 and Equ. 9, the proposed coupled feature representation is strongly dependent on two parameters: the maximum expansion power L and the balance parameter α. The influence of these two parameters will be discussed using a 10-fold cross-validation strategy with a linear SVM classifier. Here, classification accuracy was used to validate the advantage of the coupled feature representation.

Fig. 2 shows the performance of different L (i.e., $L = 2$ to 6) with the Pearson correlation coefficient. It was clear that the linear SVM classifier on coupled representation always outperformed that built on the original representation for almost every data set; the exception, D4, will be discussed in the next experiment. It was found that $L = 4$ or 5 was the empirically optimal value for capturing the global couplings of traits, which was similar to the findings of [17]. In the following experiments, we fixed L to be 4. Therefore, the hypothesis that the coupled feature representation can help improve the classification accuracy in plant breeding is accepted.

To further validate the effectiveness of our proposed coupling relationship, the influence of the balance parameter was verified by classification performance. As shown in Fig. 3, our proposed coupled representation almost always outperformed the original representation. However, for D1, the coupled representation had the worst performance while $\alpha = 0$, and this was even worse than the original representation; that is, only using the Kendall rank correlation coefficient could not capture the coupling relationships between traits.



(a) D1 (b) D2

(c) D3 (d) D4

FIGURE 2. **The performance of different *L*.**

Similarly, for D4, the coupled representation had the worst performance while $\alpha = 1$, and this was also worse than the original representation; this means that only using Pearson correlation coefficient also could not capture the coupling relationships. Consequently, the coupled representation performed worse than the original representation, as shown in Fig. 2d. It was clear that integrating the Pearson correlation coefficient and Kendall rank correlation coefficient always outperformed only using the linear correlation or the ordinal relationship. The value of the balance parameter α that is believed to reflect the best performance of different data sets may differ; this is because breeders of different trials have diverse breeding strategies. Therefore, we fixed α from 0 to 1 and reported the best results in the following experiments.

B. PERFORMANCE OF EVALUATION ALGORITHMS

Experiments were carried out to evaluate the performance of our proposed plant breeding evaluation algorithm, compared with some benchmark algorithms, including classic SVM, the Lasso algorithm, and the ElasticNet method, with both the original presentation and the coupled trait representation ($-O$ and $-C$, respectively). These comparative experiments were designed to evaluate the effect of coupled trait representation, feature selection, and the RFE algorithm. The comparison of different representations of the same method showed the impact of missing the coupled data representation component on the evaluation performance while the comparison of the same representations of different method indicated the effectiveness of RFE feature selection method. We use 10-fold cross-validation to verify their performance on plant breeding data sets.

Table 3 reports the accuracy and standard deviation results of the four approaches with the two data representations for four data sets. The bold cells indicate the best result for that data set. Larger values reflect better accuracy, and smaller standard deviations indicate better performance, because of the stronger stability of the value. It was obvious that, for each pair of the same approach for different data representations, the coupled representation outperformed the original representation with regard to both accuracy and standard deviation for almost all data sets, except for the standard deviation of Lasso-C and ElasticNet-C in D4 (highlighted by italics). The accuracy was improved by at most 7.0% and the average improvement was 3.6%; the standard deviation was reduced at most by 63%, whereas the average reduction

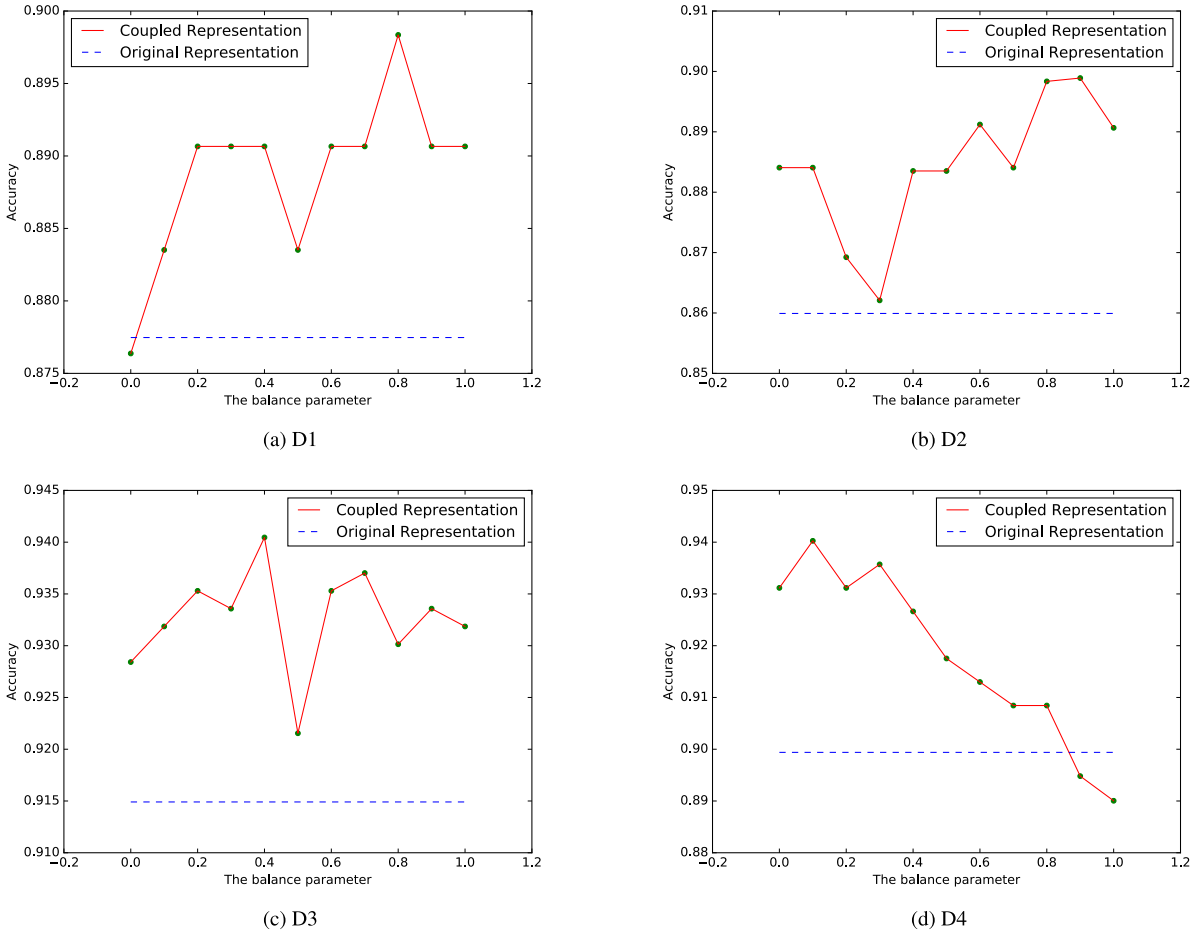


FIGURE 3. The performance of different α.

TABLE 3. Accuracy performance with ± sample standard deviation.

	D1	D2	D3	D4	Avg
SVM-O	0.877±0.159	0.860±0.165	0.915±0.018	0.899±0.099	0.888
SVM-C	0.898±0.096	0.899±0.116	0.940±**0.011**	0.940±0.068	0.919
Lasso-O	0.885±0.150	0.874±0.141	0.934±0.019	0.922±0.064	0.904
Lasso-C	0.906±0.090	0.907±0.120	0.954±0.016	0.935±*0.067*	0.926
ElasticNet-O	0.877±0.169	0.860±0.165	0.932±0.020	0.935±**0.041**	0.901
ElasticNet-C	0.920±0.075	0.920±0.081	0.954±0.016	**0.959**±*0.043*	0.938
RFE-O	0.885±0.172	0.867±0.169	0.934±0.019	0.926±0.053	0.903
RFE-C	**0.927±0.064**	**0.928±0.078**	**0.960±0.011**	0.950±0.043	**0.941**

was 39% because of the coupled trait representation. Therefore, the coupled representation can lead to accuracy and stability improvement of plant breeding evaluation algorithms.

None of the three feature selection approaches (i.e., Lasso, ElasticNet, and RFE) effectively improved the accuracy or standard deviation of SVM for the original representation. However, for coupled representation, ElasticNet-C and RFE-C improved accuracy by about 2.1% and 2.4%, respectively, and reduced standard deviation about 26% and 33%, respectively, on average, compared with SVM-C. This finding demonstrates the effectiveness of feature selection. However, Lasso-O performed similarly to SVM-O for almost all the data sets; this means that the L1 regularizer could not capture the sparse feature of coupled traits. This findings was further verified by the information included in Table 4; the selected feature numbers of Lasso-C were very different from

TABLE 4. Selected feature numbers.

	D1	D2	D3	D4
Lasso-C	22	12	18	16
ElasticNet-C	11	12	10	8
RFE-C	13	10	11	9

those of ElasticNet-C and RFE-C, whereas the latter two were very similar. Moreover, RFE-C outperformed ElasticNet-C with regard to both accuracy and stability for most data sets. In addition, as a nonparametric method, RFE is more suitable for plant breeding evaluation, because of the transformation ability between different trials.

Overall, the effectiveness of our proposed plant breeding evaluation framework was verified because RFE-C improved both accuracy and stability compared with the baseline SVM-O, with about 6.0% accuracy improvement and 56% standard deviation reduction.

V. CONCLUSION

In this paper, our computer-aided decision-making solution in plant breeding evaluation was introduced to adapt the high-throughput demand and insufficient experience of breeders in modern large-scale field breeding experiments, including a coupled feature representation scheme that integrates both linear correlation and ordinal relationship via a unified form of intra-coupled and inter-coupled relationships, and a plant breeding evaluation framework and algorithm. Experiments on several real breeding data sets were conducted to analyze the effect of preset parameters and show the effectiveness of our proposed coupled representation and plant breeding evaluation algorithm.

The contribution of this paper can be summarized as:

- A framework including data preprocessing, coupled data representation, feature selection, prediction model construction, and assisted decision-making was proposed to solve problems associated with typical plant breeding evaluation.
- A unified coupled representation scheme that integrates the intra-coupled and inter-coupled interactions was proposed to capture the interdependence among quantitative traits by addressing coupling context and coupling weights.
- A hybrid scheme of the linear correlation and ordinal relation was proposed to express the coupling relationship with a preset parameter that balances their influence in plant breeding evaluation.
- The effectiveness of the coupled representation and plant breeding evaluation algorithm was demonstrated via experiments on real plant breeding data sets that compared our approach with several benchmark approaches.

REFERENCES

[1] D. Tilman, C. Balzer, J. Hill, and B. L. Befort, "Global food demand and the sustainable intensification of agriculture," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 50, pp. 20260–20264, 2011.

[2] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, "A review on the practice of big data analysis in agriculture," *Comput. Electron. Agricult.*, vol. 143, pp. 23–37, Dec. 2017.

[3] P. L. Pingali, "Green revolution: Impacts, limits, and the path ahead," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 31, pp. 12302–12308, Jul. 2012.

[4] M. Tester and P. Langridge, "Breeding technologies to increase crop production in a changing world," *Science*, vol. 327, no. 5967, pp. 818–822, Feb. 2010.

[5] A. Rasheed, Y. Hao, X. Xia, A. Khan, Y. Xu, R. K. Varshney, and Z. He, "Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives," *Mol. Plant*, vol. 10, no. 8, pp. 1047–1064, Aug. 2017.

[6] R. K. Varshney, P. Sinha, V. K. Singh, A. Kumar, Q. Zhang, and J. L. Bennetzen, "5Gs for crop genetic improvement," *Current Opinion Plant Biol.*, vol. 56, pp. 190–196, 2020.

[7] J. L. Araus and J. E. Cairns, "Field high-throughput phenotyping: The new crop breeding frontier," *Trends Plant Sci.*, vol. 19, no. 1, pp. 52–61, Jan. 2014.

[8] Y.-Y. Han, K.-Y. Wang, Z.-Q. Liu, Q. Zhang, S.-H. Pan, X.-Y. Zhao, and S.-F. Wang, "A crop trait information acquisition system with multitag-based identification technologies for breeding precision management," *Comput. Electron. Agricult.*, vol. 135, pp. 71–80, Apr. 2017.

[9] L. Han, G. Yang, H. Yang, B. Xu, Z. Li, and X. Yang, "Clustering field-based maize phenotyping of plant-height growth and canopy spectral dynamics using a UAV remote-sensing approach," *Frontiers Plant Sci.*, vol. 9, p. 1638, Nov. 2018.

[10] Y. Li, J. Jia, L. Zhang, A. M. Khattak, S. Sun, W. Gao, and M. Wang, "Soybean seed counting based on pod image using two-column convolution neural network," *IEEE Access*, vol. 7, pp. 64177–64185, 2019.

[11] R. Khan, I. Ali, M. Zakarya, M. Ahmad, M. Imran, and M. Shoaib, "Technology-assisted decision support system for efficient water utilization: A real-time testbed for irrigation using wireless sensor networks," *IEEE Access*, vol. 6, pp. 25686–25697, 2018.

[12] Y.-Y. Han, K.-Y. Wang, Z.-Q. Liu, S.-H. Pan, X.-Y. Zhao, and S.-F. Wang, "Golden seed breeding cloud platform for the management of crop breeding material and genealogical tracking," *Comput. Electron. Agricult.*, vol. 152, pp. 206–214, Sep. 2018.

[13] A. R. Hallauer, "History, contribution, and future of quantitative genetics in plant breeding: Lessons from maize," *Crop Sci.*, vol. 47, pp. S-4–S-19, Dec. 2007.

[14] X. Zhao, Z. Liu, F. Dan, and K. Wang, "Plant breeding evaluation with rank entropy-based decision tree," *IFAC-PapersOnLine*, vol. 49, no. 16, pp. 336–340, 2016.

[15] Y. Cao, Y. Jiang, H. Gao, H. Chen, X. Fang, H. Mu, and F. Tao, "Development of a model for quality evaluation of litchi fruit," *Comput. Electron. Agricult.*, vol. 106, pp. 49–55, Aug. 2014.

[16] Z. Zhai, J. F. Martínez, V. Beltran, and N. L. Martínez, "Decision support systems for agriculture 4.0: Survey and challenges," *Comput. Electron. Agricult.*, vol. 170, Mar. 2020, Art. no. 105256.

[17] C. Wang, Z. She, and L. Cao, "Coupled attribute analysis on numerical data," in *Proc. IJCAI*, 2013, pp. 1–7.

[18] L. Cao, "Coupling learning of complex interactions," *Inf. Process. Manage.*, vol. 51, no. 2, pp. 167–186, Mar. 2015.

[19] C. Wang, C. Chi, W. Zhou, and R. K. Wong, "Coupled interdependent attribute analysis on mixed data," in *Proc. 29th AAAI Conf. Artif. Intell.*, Austin, TX, USA, Jan. 2015, pp. 1861–1867.

[20] W. Cao and L. Cao, "Financial crisis forecasting via coupled market state analysis," *IEEE Intell. Syst.*, vol. 30, no. 2, pp. 18–25, Mar. 2015.

[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

XIANGYU ZHAO received the B.S. degree in automation from Beijing Jiaotong University, China, in 2007, and the M.S. degree in software engineering and the Ph.D. degree in computer software and theory from the Beijing Institute of Technology, China, in 2009 and 2014, respectively. He is currently an Assistant Professor with the Beijing Research Center for Information Technology in Agriculture. His current research interests include data mining, agricultural information, and agricultural intelligent systems.

YANYUN HAN received the Ph.D. degree from the School of Information Science and Technology, Beijing Forestry University, in 2015. She is currently a Research Assistant with the Beijing Research Center for Information Technology in Agriculture, Beijing, China. Her research interests include the agriculture information technology, seed industry informationization, and crop breeding data management systems.

SHOUHUI PAN received the Ph.D. degree in management science and engineering from the Beihang University of China, in 2012. He is currently an Associate Professor with the National Engineering Research Center for Information Technology in Agriculture, China. His current research interests include data mining, intelligent information processing, and management information systems.

ZHONGQIANG LIU received the Ph.D. degree in agricultural information technology from China Agricultural University, China, in 2016. He is currently an Associate Professor with the Beijing Research Center for Information Technology in Agriculture. His current research interests include precision agriculture and agricultural information platform.

KAIYI WANG received the Ph.D. degree from the School of Computer Science, Beijing Industry University, Beijing, China. He is currently a Professor with the Beijing Research Center for Information Technology in Agriculture, Beijing. His current work interests focus on plant breeding information research and application.