Breeding Science 48 : 251~ 255 (1998)

Evaluation of Soybean Plant Shape Based on Tree-Based Models

Seishi Ninomiya and Vu Nguyen-Cong

National Agriculture Research Center, Tsukuba, Ibaraki 305-8666, Japan

Summary

Plant shape, an important factor in soybean plant breeding, is currently evaluated visually by soybean plant breeders, often making judgment unstable and subjective. The purpose of our study was to create procedure for objectively evaluating soybean plant shape. Features of shape were determined by image analysis. Tree-based models based on recursive partitioning were then used to categorize shapes into three classes -- "good," "fair" and "poor" -- or two classes -- "good" and "not good." Classification results based on tree-based models demonstrated highly acceptable predictability. Although model-based performance attained approximately the same discriminatory level as conventional linear discriminant function, it had the distinct advantage of outstanding interpretability, with shape parameters in the best predictive tree-based model coinciding with those selected empirically by expert breeders.

Key Words : classification, *Glycine max* L. Merrill, plant shape, recursive partitioning, tree-based models.

Introduction

In soybean breeding, plant shape is one of the important selection factors, because shape is relating to lodging resistance, light interception, amount of yield and adaptability for machine harvesting (Ninomiya and Shigemori 1991). Conventional visual shape quality judgment by expert breeders may involve unstable or subjective choice, and attempts have been made to automate judgment combining classification models with shape parameters extracted by image analysis. In two successful cases, Ninomiya and Shigemori (1991) applied a linear discriminant function to classification and Ambuel *et al.* (1997) applied fuzzy logic. Both, however, had drawbacks. The linear discriminant function approach assumes classes to be normally distributed with equal covariance matrices. This assumption leaves a limitation to practical applications as the normality is not always expected. Moreover, in the linear discriminant function, it was not intuitively easy to interpret the effect of shape parameters on classification and to understand the relationship between model classification and human judgement. In the fuzzy logic approach, although it was easy to interpret individual shape parameters in the model, once rules were fixed, rule-making combining appropriate parameters and adjusting fuzzy ranges was *ad hoc*, difficult and time-consuming.

In our study, we worked to overcome the above drawbacks by applying tree-based models (Breiman *et al.*, 1984).

Received September 12, 1997. Accepted June 10, 1998.

Materials and Methods

Experimental data

Three expert soybean breeders evaluated visually the shape quality of 875 whole plant images of soybean provided by Ninomiya and Shigemori (1991). Their judgments were classified into 1 for "poor," 2 for "fair" or 3 for "good." Because uneven judgments resulted revealed the unavoidable subjectivity of the human factor, we used only the 325 shape samples with coincident judgments made by the three breeders in this study. Results yielded 166 shapes judged class 1, 93 judged class 2 and 66 judged class 3. Focusing on "good" shape, however, judgment corresponding to "fair" and "poor" cases could be combined into one class, posing a problem with dichotomous classification. Twenty shape parameters, defined by Ninomiya and Shigemori (1991), with binary image analysis were used as input variables to develop classification tree models (Fig. 1).

Classification trees

Classification is a learning problem in which a value of output variable y to be predicted falls into one of K classes, $y \in \{1, 2,..., K\}$. Prediction is based on p measurements of parameters or input variables $x = \{x_1, ..., x_p\}$. The rule for predicting the class membership of observation x is constructed by a learning algorithm from training set T of n previously observed cases, $T=\{(x_i,y_i)|\ i=1,2,...,n\}$, for which the joint values of both input and output variables are given. Such learning algorithms include linear discriminant analysis, multiple logistic regression, nearest neighbor methods, artificial neural networks, and tree-based methods (Ripley 1996, Bishop 1995).

Tree-based modeling (Breiman *et al.* 1984) is a statistical technique for describing knowledge and making decisions. It can produce a classification tree. To predict the class membership of a new observation, the process starts at the root node and descends to either a left or right node, depending on the discrimination value of the currently visited node. This is done recur-

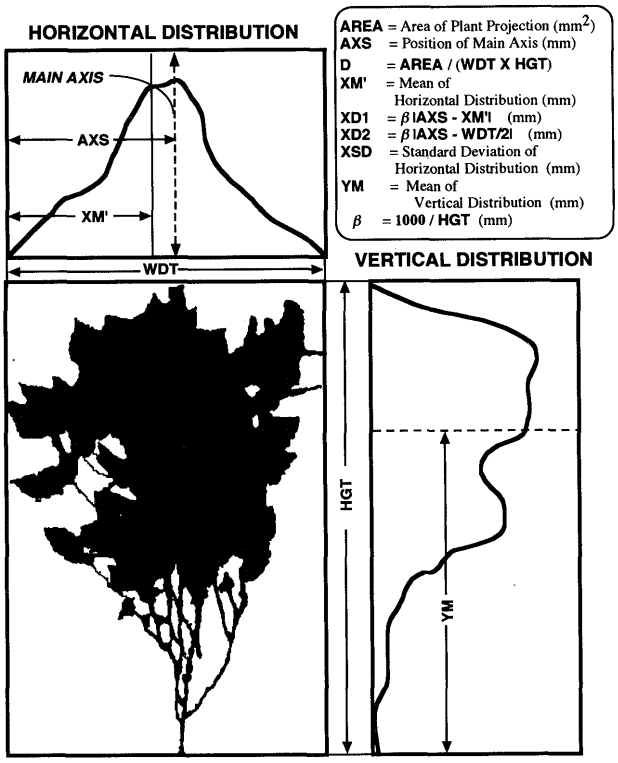

Fig. 1. Shape parameters given by Ninomiya and Shigemori (1991). Only parameters discussed in this article are shown. β is a coefficient to normalize parameters to adjust the height of a plant to 1,000 mm. Note that XD1, XD2, XSD, and YM are normalized values. XM' differs from the original XM (Ninomiya and Shigemori 1991) because XM is normalized but XM' is not.

sively until a terminal node (leaf) is reached. Leaf values are then used to make a decision. The tree representation, described below, thus provides powerful graphic visualization, compared with many other classification methods.

Partitioning starts with single region R_0 containing entire training set T. At each step in partitioning, each "parent" region is optimally split into two "daughter" subregions, R_{left} and R_{right}, which are then themselves split, thereby increasing the number of regions. To determine how to split training set T into subregions, a measure of impurity is defined so that a region's impurity is largest when all K classes are equally distributed within it, and smallest when each region contains only one class. A particular split is then selected to produce the greatest decrease in average impurity. Recursive partitioning continues until no further split is found, to improve the purity of terminal nodes, or until the number of observations reaching each terminal node becomes too small (≤ 5 by default in S-PLUS).

Instead of using a measure of impurity, the tree construction process in S-PLUS uses a deviance function for the tree (Clark and Pregibon, 1992). It is defined as

$$D=\sum_i D_i, \qquad D_i= -2\sum_i n_{ik}\log(p_{ik})$$

where n_{ik} is the number of observations of k^{th} class in i^{th} terminal node and p_{ik} is the multinomial distribution parameter estimated from the node proportion, $n_{ik}/\sum_k n_{ik}$.

In the tree-growing process, the variable used for each partitioning was selected from the subset member and the value for the partitioning was one of the mid values of all the possible adjacent observations. The selections of the variable and the partition point were based on the deviance evaluated at each partition.

The classification tree is a collection of decision rules determined by recursive partitioning. In practice, class distributions usually overlap each other, possibly preventing a complete partition. We thus adopt a tree with the minimum error rate, *i.e.*, the proportion of the training sample being misclassified, defined as

$$Error\ rate = \frac{number\ of\ misclassifications}{number\ of\ observations}$$

The tree-growing process discussed above may produce an overly large tree that overfits a model to the training set, leading to a situation in which randomized training-set characteristics, rather than general training-set features, are modeled. Thus, the next step is to reduce the tree size by pruning. The pruning is performed by snipping off the nodes under the least important node not being a leaf in the tree. The importance of the i^{th} node, denoted by α_i, is defined as

$$\alpha_i = \frac{m_i - \sum_j m_j^i}{n_i - 1}$$

where m_i ,m_j^i and n_i are the number of misclassification at the i^{th} node, the number of misclassification at the j^{th} leaf under the i^{th} node and the total number of leaves under the i^{th} node, respectively. When more than one nodes are least important with the equivalent α values at the same pruning stage, the nodes under those nodes are all snipped at once. The next pruning is performed on the adjacently pruned tree, recalculating the α_i values. This pruning process is repeated until the root node becomes least important. For each pruning process, the error rate for the misclassification and the least important node value (α) denoted by μ_l (l = 0, 1, 2, 3,.....) for the l^{th} pruning process, are stored in a sequence.

In this pruning process, the more pruned, the larger the error rate generally. The error rate, however, does not indicate the predictability of the model and the most appropriate tree size for the best predictability should be found among the sequence of the trees pruned with the different μ_l values. In this study, a k-fold cross validation (Stone, 1974) was used to find the reasonable tree size and the final tree chosen is a subtree having the smallest estimated prediction error. Selection of a k value in k-fold cross validation is usually recommended as 10 (Efron and Tibshirani 1993, Ripley 1995). Merler and Furlanello (1997) recently introduced the 632+ bootstrap method (Efron and Tibshirani, 1995) in tree-

Table 1. Good predictive models in dichotomous classification with combined "poor" and "fair" classes

Model[1)]	Input variables[2)]	Error rate of original tree[3)]	Error rate of pruned tree[4)]	Leaf number of pruned tree[4)]	CV for pruned tree
1	D, XM, XSD	.06(19/325)	.08(25/325)	8	.14(46/325)
2	WDT, D, XD1, XSD	.04(13/325)	.07(22/325)	8	.13(43/325)
3	XD1, WDT, D, YM, XSD	.03(10/325)	.03(10/325)	15	.13(42/325)
4	XD1,WDT, D, XD2, XSD	.03(10/325)	.05(16/325)	9	.11(35/325)

[1)] The best models for the subset with one or two variables are not shown as the error rates for those were considerably high.

[2)] The definitions for the variables are shown in Fig. 1.

[3)] Error rate for the best fitted models before pruning (the step 2 of **Appendix**).

[4)] Error rate for the tree pruned with μ_l selected based on CV (the step 3 (1) of **Appendix**).

See text for further details.

based model selection. For 10-fold cross validation, the data set is randomly split into 10 roughly equal parts, of which 9 are then used to grow the tree and the rest one is used to assess the misclassification. This is done 10 times and the estimated prediction error calculated over the 10 runs is defined as,

$$CV = \frac{\sum_{i=1}^{10} number\ of\ misclassifications\ in\ i^{th}\ part}{number\ of\ observations(n)}$$

In the cross validation, the α_i values for the grown trees are calculated and the trees are pruned at the stems below the node whose importance is equal or less than μ_l. Then, the CV values are compared among the trees pruned with different μ_l values to find the best tree size.

Tree construction used here is an improved version of the technique described by Clark and Pregibon (1992) in S-PLUS (Venables and Ripley, 1994). The data set involved a large number of input variables (20), some highly mutually correlated, requiring that variables be selected before proceeding with the tree-based modeling to reduce data redundancy and attain a simpler tree model to interpret. To prevent the time-consuming work entailed in investigating all possible variable subsets, we first exhaustively searched data sets including only one input variable, then continued the work with sets having two input variables, and so on, until a tree-based model with good predictive ability was obtained. The maximum number of input variables was fixed at five here, because $\binom{20}{5}$=15504 models must be investigated and we did not have the computational capacity required to process more than this number.

The above process is summarized in **Appendix**.

Results and Discussion

Our first experiment considered dichotomous classification. In the four recommended models with up to five input variables (Table 1), the most complicated (Model 3) had 15 leaves and its prediction error (CV) did not outperform other models. Because the less complicated model is recommended to avoid overfitting among the models with almost the same error rates, we eliminated it from consideration. In the remaining models, Model 4 had the smallest prediction error, even though it contained one more leaf than Models 1 and 2, and was chosen as the final model (Fig. 2 and Table 2). Variables XD1 and D were used to build the tree's upper branches, indicating importance in evaluating soybean plant shape. These two variables are the same as those selected by expert breeders in fuzzy logic classification for the same data (Ambuel *et al.* 1997). The tree-based model thus provides a tool for analyzing the visual judgment of experts-- something not possible using the conventional linear discriminant function.

A tree-based model for the 3-class problem was determined based on the best subset of input variables in dichotomous classification (Model 4, Table 1). The resulting 3-class tree model (Fig. 3, Table 3) contained 12 leaves, and was more complicated than the tree model used in the dichotomous case (Fig. 2). Combining classes 1 and 2 in Table 3 makes it clear that this model is similar to Model 4 (Table 2), showing that the 3-class tree model is not necessary for interpreting shape variables and evaluating soybean plants.

For comparison, we applied conventional linear discriminant analysis (Venables and Ripley 1994) to the same data. The most suitable model found included five input variables, *i.e.*, HGT, AREA, XD1, YM and XSD. Adding more variables did not significantly improve fitting; the error rate for this model was 0.095 and the prediction error 0.12, suggesting little loss in accuracy, compared with the best tree-based model.

The tree-based model's predictability was high

Table 2. Model 4 classification table

Observed Class	*Predicted Class* 1+2	3
1+2	253	6
3	10	56

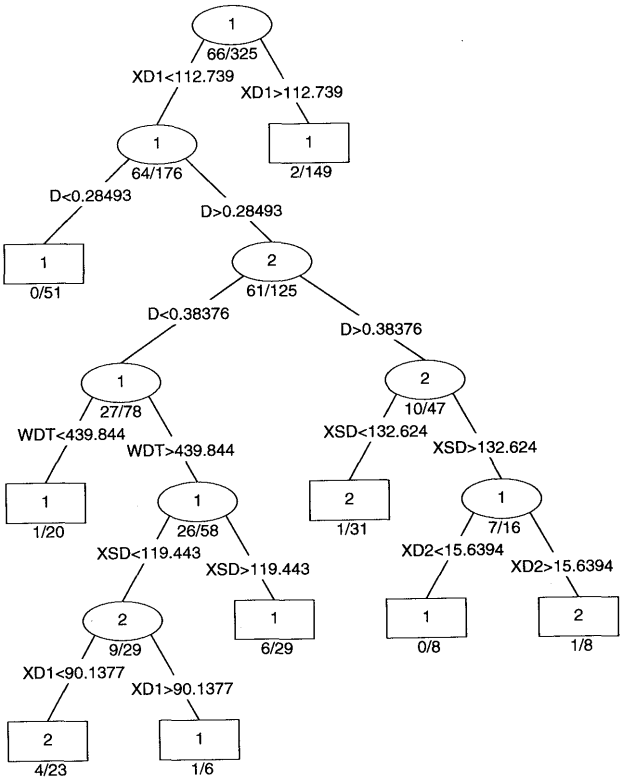

Fig. 2. Model 4 tree representation after pruning. Class 1 combines "poor" and "fair" judgments. Class 2 is for "good" observations. Inner nodes are denoted by ovals and terminal nodes (leaves) by rectangles, with the predicted value of the output variable centered in the node. The number under each node indicates the error rate, *m/n*, where *m* and *n* are the number of misclassification and the number of the observations at the node, respectively.

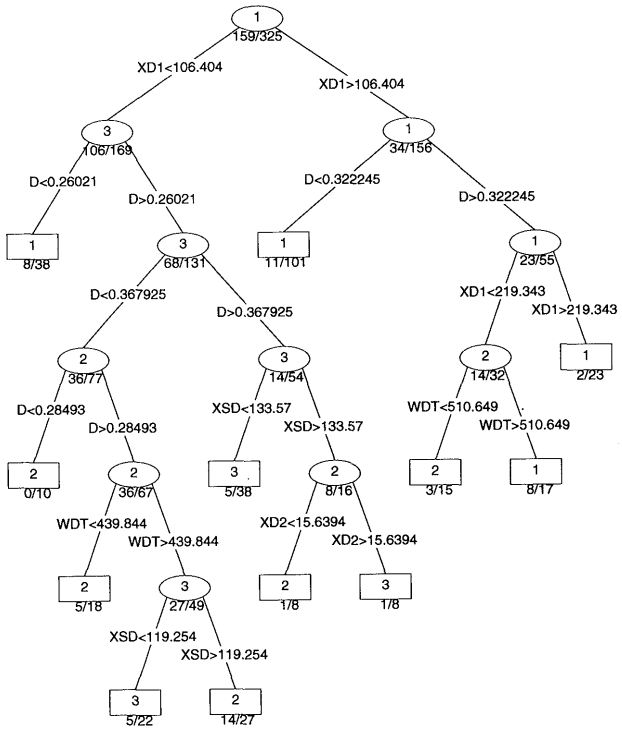

Fig. 3. Three class classification problem tree representation after pruning. The input variables are the same as the Model 4 variables. See text and the legend of Figure 2 for details.

Table 3. Three class tree model classification table

Observed Class	*Predicted Class*		
	1	2	3
1	150	16	0
2	27	55	11
3	2	7	57

enough to substitute for visual classification. Despite the small gain in predictive accuracy in evaluating soybean plant shape, compared to conventional linear discriminant analysis, tree-based models have the distinct advantage of classification results that are readily explainable and justifiable. Interpretability is important because breeders are at ease with predictions resulting from a model whose operation they can readily understand. Tree-based models were originally designed to resolve data complexity and high dimensionality, and a variable subset may have to be selected before modeling to reduce data redundancy and attain a tree model simpler to interpret. It should be noted that our limited computational capacity prevented us from investigating all possible combinations of input variables, and that our "best" model may not yet be "the ultimate best".

Acknowledgment

This work was financially supported by the Japanese Science and Technology Agency (STA).

Literature Cited

Ambuel, J., S. Ninomiya and N. Takahashi (1997) Fuzzy logic soybean plant shape evaluation. Breed. Sci. 47:253-257.

Bishop, C. M. (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford, pp 504.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984) Classification and Regression Trees. Wadsworth, Belmont, CA, pp 358.

Clark, L. A. and D. Pregibon (1992) Tree-based models. *In* "Statistical Models in S," Chambers, J. M. and T. J. Hastie (eds.), Chapman & Hall, New York, 377-419.

Efron, B. and R. Tibshirani (1993) An Introduction to the Bootstrap. Chapman & Hall, New York, pp 436.

——, and ——(1995) Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Tech. Rep. 176, Stanford Univ.

Merler, S. and C. Furlanello (1997) Selection of tree-based classifiers with the bootstrap 632+ rule. Biometr. J. 39:1-

14.

Ninomiya, S. and I. Shigemori (1991) Quantitative evaluation of soybean (*Glycine max* L. Merrill) plant shape by image analysis. Jpn. J. Breed. 41:485-497.

Ripley, B. D. (1995) Statistical ideas for selecting network architectures. *In* "Neural Networks: Artificial Intelligence and Industrial Applications," Kappen, B. and S. Gielen (eds.), Springer-Verlag, London, 183-190.

—— (1996) Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, pp 403.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). J. Roy. Statist. Soc. Ser. B 36:111-147.

Venables, W. N. and Ripley, B. D. (1994) Modern Applied Statistics with S-Plus. New York: Springer-Verlag, pp 462.

Appendix

Process to obtain recommended tree-based models

1. A tree for each of the variable subsets of ${}_{20}C_P$ (P=1,2, 3, 4, 5) combinations from twenty shape variables (Ninomiya and Shigemori 1991) is grown based on the deviance D and the minimum number of a terminal node member ($\leqq 5$).
2. The tree that has the minimum fitted error rate in each variable subset size (P=1, 2, 3, 4, 5) is left for the step 3 as an original candidate tree. When the error rates are equivalent, more than one candidate for each size can be left.
3. For each tree left in the step 2, a nested sequence of subtrees is determined by recursively snipping off the least important splits and the most predictable tree is chosen from the sequence of the subtrees based on the cross validation, as follows.

(1) The importance of the i^{th} node not being a leaf in the tree, denoted by α_i, is calculated for each node and the pruning is done at the stems below the least important node (s) with the least α value (s).

(2) The step 3-(1) is recursively done for the pruned tree, until the root node becomes least important so that no more pruning is possible. This recursive process produces a sequence of the least important node values for each pruning denoted by $\mu_0\ \mu_1\ \mu_2\ \mu_3$, ... and the corresponding pruned trees.

(3) The 10-fold cross validation is applied to estimate the CV for each of the pruned trees in the sequence. The grown tree in the cross validation is pruned based on the μ_l values.

3. The best μ_l with the least CV can be found in the step 3-(3) for each original tree left in the step 2. Then, the pruned tree at the step 3-(1) corresponding to the best μ_l value is selected as a recommended model for each original tree.