LARS Information Note 040869
Purdue University

An Automated Method of Choosing Training Fields

by

Danielle R. Bernstein

When a flight line of data is to be analyzed using LARSYSAA, a set of representative training fields must be chosen. This set must be able not only to recognize (classify) itself but to recognize other fields (test fields) as well. Therefore it is important that an automatic method be devised to give the user an initial set of training fields. If necessary, this set can be refined in subsequent classifications to make it more representative of the test fields.

An experiment was conducted to evaluate such a method. A program was written (PPS - Serial No. DA 0013) to choose training fields within each class with probability proportional to size (PPS).

The fields are chosen in the following manner. The number of samples in each field is calculated (Mi) and a sum total is kept (ΣMi). Let the following table be an example.

Unit	Size Mi	ΣMi	Assigned Range
1	3	3	1- 3
2	1	4	4
3	11	15	5-15
4	6	21	16-21
5	4	25	21-25

A random number is chosen between 1 and 25. Suppose it is 19. In the sum, number 19 falls in unit 4 which covers numbers 16 to 21 inclusive. With this method of drawing, the probability that any unit is selected is proportional to the size of the unit. Sampling is done without replacement and therefore unit 4 cannot be picked again. In addition, the program will optionally scale the training fields to a desired number of samples (through manipulation of the line and column intervals) so that no one field will have an undue influence on the statistics of the training classes.

With this method, notice that:

1) The number of fields and the range of their sample size are selected by the user.
2) The training fields are chosen from all the fields which can be outlined and for which there is ground truth, i.e. from all fields to be used as test fields No attempt is made to delete the nonuniform or atypical fields nor are the boundaries changed so that only the center part is used.
3) The fields are chosen from each class to be separated (not subclasses) and the histograms of the training classes may not be unimodal. However, this results in fewer training classes and faster classification time.

Two runs (September, 66-C3 and C4) were each analyzed twice (see computer output in file) in this manner and compared with their analysis done previously in the conventional manner (see Information Note D022469).

C3 - Five classes were to be separated; soybeans, corn, stubble, forage and water. Water consisted of only one segment which was not picked by PPS. The following number of fields were chosen initially.

Class	Training	Test
Soybeans	4	10
Corn	5	13
Stubble	4	10
Forage	4	9
Water	1	1

Histograms revealed that although none of the classes were unimodal, the stubble and forage classes were so bimodal that some adjustment had to be made. One field was removed from the stubble class and forage was divided into two subclasses; FRG1, and FRG2 made up of two fields each. The feature selection processor showed that there was poor separation between soybeans and corn but adequate-to-good separation between the rest of the fields. Features 1, 6, 9; 10 were chosen.

The classification results for training classes were 80.7% for overall performance compared with 91.9% for the previously documented classification. This is to be expected as the training classes were not unimodal and not picked for their uniformity. The test class results were 72.4%. This compares favorably with 71.0% obtained in the previous classification. This would tend to indicate that the training fields, when picked at random, do not result in worse classification of the test set than when carefully chosen through several iterations. The latter (previously documented) classification took three or more iterations to arrive at the final set of training samples. Another classification of C3 using five

classes gave similar results - 81.8% for training fields and 69.6% for test fields.

C4 - C4 was also classified twice in this manner. Only 4 classes (and no subclasses) were needed; soybeans, corn, stubble and forage. The original classification resulted in 91.0% overall recognition of training classes and 73.1% for test classes. The classification, with training fields chosen with PPS, used channels 1, 6, 10, 12. The classification resulted in 80.6% overall performance for the training classes and 69.6% for test classes. For the second classification of C4, the same classes and features 1, 6, 9, 10 were used. Training class recognition was 85.2% and test class recognition was 68.0%.

Although test field recognition was a little lower in C4, this decrease has to be judged in the light of the number of classifications performed to arrive at the final results. When classifying aircraft data in the conventional manner, the user is essentially working with his training fields. He refines his set and tries to improve his training fields classification performance. Only when he is satisfied with this performance, does he really concentrate on his test fields. However, it is test field accuracy that is important and which must be improved.

LARSYSAA users are encouraged to employ this program to obtain at least a starting training set, "Mild" multimodality can be ignored. In this way, fewer subclasses will be used and higher accuracy with fewer iterations will result.