Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

June 29 - July 1, 1976

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
76CH1103-1 MPRSD

SELECTING CLASS WEIGHTS TO MINIMIZE CLASSIFICATION BIAS IN ACREAGE ESTIMATION*

W. M. Belcher and T. C. Minter

Lockheed Electronics Company, Inc./Aerospace Systems Division, Houston, Texas

I. ABSTRACT

Classification of multispectral data by the use of a maximum likelihood classifier is dependent upon knowing in advance a set of prior probabilities. Therefore, the selection of an optimal set of prior probabilities is critical to the estimation of proportions for each class. In the proposed procedure, a function is minimized to yield a set of optimal prior probabilities for a specific data set. Classification results using optimal, actual, and default (equal prior probabilities for each class) values are compared.

II. INTRODUCTION

Often, when classifying multispectral data, the proportion of each crop (or class) in a given area is estimated by counting the number of pixels (picture elements) classified into each class and normalizing this with the total number of pixels in the area. This proportion estimate is sometimes biased by the Bayes decision rule used in the classification. The decision rule for assigning a sample X (a pixel from a multispectral image) to class i is

$$k_i p(X/i) \geq k_j p(X/j) \quad \text{for } j = 1,\cdots,m \quad (1)$$
$$j \neq i$$

where k_i, $i = 1,\cdots,m$, is the class weight, which is usually taken to be an estimate of the class prior probability. The class conditional density function is denoted by $p(X/i)$, $i = 1,\cdots,m$, where m is the number of classes. The sample X is assumed to be a random p-dimensional measurement vector drawn independently from the sample.

This paper presents preliminary results of experimentation being done to select optimal class weights for use with the maximum likelihood classifier. These weights will be optimal in the sense that the bias will be minimized in the proportion estimate obtained from the classification results by sample counting.

Results will be presented for the special case where $m = 2$ (i.e., wheat and nonwheat). The use of optimal class weights will be compared with the use of equal class weights and the use of the true proportions for class weights.

III. MEAN SQUARE ESTIMATION OF OPTIMUM VALUES FOR THE CLASS WEIGHTS

In this section an analytical procedure is presented for estimating optimum values for the class weights $(k_i,\ i = 1,\cdots,m)$ based on historical values for the class (crop) proportions. Mean square error criteria will be minimized to obtain optimal values for k_i, $i = 1,\cdots,m$. The values for k_i, $i = 1,\cdots,m$, obtained from minimizing these criteria are optimal in the sense that the bias in the proportion estimate obtained from the classification results is minimum.

Let the vector $Q_t = \left\{q_1^t, q_2^t, \cdots, q_m^t\right\}$ be the historical estimates of proportions for the m classes (crops) in the area during the t*th* year past. Also, it is assumed that historical data on class proportions are available for n past years; i.e., $t = 1,\cdots,n$. Each class is assumed to be normally distributed with class conditional probability density function $p(X/j)$, $j = 1,\cdots,m$. The mean vectors $\hat{\mu}_j$, $j = 1,\cdots,m$, and covariance matrices $\hat{\Sigma}_j$, $j = 1,\cdots,m$, for $p(X/j)$, $j = 1,\cdots,m$, are assumed to have been estimated from

*The material for this paper was developed under Contract NAS 9-12200 for the Earth Observations Division, Science and Applications Directorate, Lyndon B. Johnson Space Center, National Aeronautics and Space Administration, Houston, Texas.

the current year's multispectral imagery data.

Based on historical class proportions from the t*th* year, Q_t, the overall mixture density is

$$p_t(X) = \sum_{j=1}^{m} q_j^t \, p(X/j) \qquad (2)$$

For a given k_i, $i = 1,\cdots,m$, and $p_t(X)$ (as defined using proportions Q_t from year t), the proportion of samples classified as class i is

$$Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m) = \int_{R_i} p_t(X)dX \qquad (3)$$

where R_i is the Bayes region for class i;

$$R_i = \{X | k_i p(X/i) \geq k_\ell p(X/\ell)\} \qquad (4)$$

$$\text{for } \ell = 1,\cdots,m$$
$$\ell \neq i$$

For the i*th* class in the t*th* year the error in the proportion estimate provided by $Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m)$ is

$$Error_i^t = q_i^t - Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m) \qquad (5)$$

Based on historical proportions Q_t, $t = 1,\cdots,n$, for n years past and considering m classes, the following mean square error criterion is proposed to evaluate the error in the proportion estimate obtained from the classification results for a particular set of class weights k_i, $i = 1,\cdots,m$;

$$J = \frac{1}{nm} \sum_{t=1}^{n} \sum_{i=1}^{m} \left[q_i^t - Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m) \right]^2 W_t \qquad (6)$$

The year weights W_t, $t = 1,\cdots,n$, may be selected in order to give more weight to more recent information on crop proportions and therefore adapt for trends.

The minimization of these criteria with respect to k_i, $i = 1,\cdots,m$, is considered next. The class weights will be subject to the following constraints:

$$k_i \geq 0 \quad ; \quad i = 1,\cdots,m \qquad (7)$$

$$\sum_{i=1}^{m} k_i = 1 \qquad (8)$$

Using a Lagrange multiplier, the second constraint (eq. 8) may be enforced by modifying equation (6) to

$$J = \frac{1}{nm} \sum_{t=1}^{n} \sum_{i=1}^{m} \left[q_i^t - Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m) \right]^2 W_t + \lambda\left(\sum_{i=1}^{m} k_i - 1 \right) \qquad (9)$$

The optimum values for the class weights may be found by using the expression for J, $\frac{\partial J}{\partial k_i}$, $i = 1,\cdots,m$, and $\frac{\partial J}{\partial \lambda}$ (eqs. 9, 21, and 22) in a numerical optimization procedure.[1]

In the narrative that follows, the differentiation of J (eq. 9) with respect to k_i, $i = 1,\cdots,m$, and λ is discussed.

In order to differentiate J (eq. 9) with respect to k_i, $i = 1,\cdots,m$, an alternative form for $Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m)$ (eq. 3) must be obtained which can be differentiated.

The Bayes region for class i, R_i, is defined (eq. 4) as

$$R_i = \{X | k_i p(X/i) \geq k_j p(X/j)\} \qquad (10)$$

$$\text{for } j = 1,\cdots,m$$
$$j \neq i$$

The following function sgn(y) is defined:

$$sgn(y_{ij}) = \begin{cases} 1 \text{ if } y_{ij} \geq 0 \\ 0 \text{ if } y_{ij} < 0 \end{cases} \qquad (11)$$

where $y_{ij} = k_i p(X/i) - k_j p(X/j)$. (12)

The Bayes region for class i, R_i, is now redefined as

$$R_i = \left(X \left| \prod_{\substack{j=1 \\ j\neq i}}^{m} sgn(y_{ij}) > 0 \right. \right) \qquad (13)$$

Substituting $[\frac{1}{2} + \frac{1}{2}\tanh(sy_{ij})]$ for $\text{sgn}(y_{ij})$, it can be seen that

$$\left[\frac{1}{2} + \frac{1}{2}\tanh(sy_{ij})\right] \cong \begin{cases} 1 \text{ if } y_{ij} > 0 \\ 0 \text{ if } y_{ij} < 0 \end{cases} \quad (14)$$

where s is a large positive real number. (Note: When $y_{ij} = 0$, this function has the value 1/2, but this can be ignored since it would occur on a set of measure zero.)

Let $$z_{ij} = \frac{1}{2} + \frac{1}{2}\tanh(sy_{ij}) \quad (15)$$

The Bayes region for class i, R_i, can be written (using eq. 15) as

$$R_i \cong \left(X \Big| \prod_{\substack{j=1 \\ j\neq i}}^{m} z_{ij} > 0\right) \quad (16)$$

From equation (16), let

$$f_i(k_1,k_2,\cdots,k_m) = \prod_{\substack{j=1 \\ j\neq i}}^{m} z_{ij} \quad (17)$$

The expression for $Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m)$, equation (3), is now rewritten in a form that is differentiable.

$$Pr_t(X\varepsilon R_i | k_1,k_2,\cdots,k_m) = \int_{-\infty}^{\infty} p_t(X) f_i(k_1,k_2,\cdots,k_m)dX \quad (18)$$

Substituting equation (18) into equation (9), the criteria J is rewritten as

$$J = \frac{1}{nm}\sum_{t=1}^{n}\sum_{i=1}^{m}\left[q_i^t - \int_{-\infty}^{\infty} p_t(X) f_i(k_1,k_2,\cdots,k_m)dX\right]^2 W_t + \lambda\left(\sum_{i=1}^{m} k_i - 1\right) \quad (19)$$

The partial derivative of J with respect to k_i is

$$\frac{\partial J}{\partial k_i} = \frac{2}{nm}\sum_{t=1}^{n}\sum_{i=1}^{m}\left[q_i^t - \int_{-\infty}^{\infty} p_t(x) f_i(k_1,k_2,\cdots,k_m)dx\right] \cdot \int_{-\infty}^{\infty} p_t(x)\frac{s}{2}\, p(x/i) \sum_{\substack{j=1 \\ j\neq i}}^{m}\left[\left(\prod_{\substack{r=1 \\ r\neq j \\ r\neq i}}^{m} z_{ir}\right) \cdot \text{sech}^2(sy_{ij})\right]dx + \lambda \quad (20)$$

This can be expressed in terms of an expected value

$$\frac{\partial J}{\partial k_i} = \frac{2}{nm}\sum_{t=1}^{n}\sum_{i=1}^{m}\left[q_i^t - E_t\{f_i(k_1,k_2,\cdots,k_m)\}\right] \cdot E_t\left\{\frac{s}{2} p(X/i) \sum_{\substack{j=1 \\ j\neq i}}^{m}\left[\left(\prod_{\substack{r=1 \\ r\neq j \\ r\neq i}}^{m} z_{ir}\right)\text{sech}^2(sy_{ij})\right]\right\} \quad (21)$$

where expectation operator E_t is defined as $E_t(\cdot) \int_{-\infty}^{\infty} (\cdot) p_t(X)dx$.

The partial derivative of J (eq. 19) with respect to λ is

$$\frac{\partial J}{\partial \lambda} = \sum_{i=1}^{m} k_i - 1 \quad (22)$$

IV. DESCRIPTION OF EXPERIMENT

A. The Data

The procedure was tested using Landsat multispectral scanner (MSS) data from an 8- by 9.6-kilometer (5- by 6-mile) area of ground truth in Finney County, Kansas. Data consisted of four passes. However, one channel of data on the first pass was of such poor quality that it was not usable.

A set of 35 training fields was selected at random from the site conditioned on the field containing 19 pixels or more. The set was then divided into two classes, wheat and nonwheat, and clustered into three and five subclasses, respectively, thus allowing the investigation of a two-class, eight-subclass case.

Actual prior probabilities for the two classes were computed by dividing the number of acres in each class by the total number of acres in the test site.

B. Results

The results for Finney County, Kansas, are summarized in table 1. The bias in table 1 was calculated from

Bias = Estimated wheat proportion minus true wheat proportion

decrease in wheat proportion bias was slight (see table 1); i.e., approximately 2 percent.

C. Analysis of Results

The decrease in proportion bias obtained for the simulated data, when optimal class weights were used instead of equal class weights, was not observed when imagery data were used. In addition, for the data sets used (i.e., for simulated and imagery data), very little decrease in proportion bias for wheat was observed using true proportions for class weights in place of optimal class weights (1.1 percent for both data sets).

There are several possible explanations for these results. The differences observed in the results between the simulated data and the imagery data may be attributable to (1) unrepresentative training statistics (i.e., not all classes represented, the presence of boundary pixels, etc.) and (2) inappropriate estimates of the true proportions of the wheat and nonwheat subclasses (each subclass of wheat and nonwheat was given equal weight within the class for purposes of estimating optimal class weights).

The lack of differences observed in proportions estimated using true proportions for class weights, instead of the optimal class weights, can probably be attributed to having well-separated classes (the probability of correct classification for wheat was 93.6 percent and 93.3 percent for nonwheat).

Table 1. Experiment Results

True proportions	Bias observed for wheat in the simulated data using the class weights indicated below		Bias observed for wheat in the Finney County imagery data using the class weights indicated below		
	Equal $K_w = 0.5$ $K_o = 0.5$	True proportions $K_w = 0.25$ $K_o = 0.75$	Equal $K_w = 0.5$ $K_o = 0.5$	True proportions $K_w = 0.25$ $K_o = 0.75$	Optimal $K_w = 0.254$ $K_o = 0.745$
Wheat = 0.25 Nonwheat = 0.75	17.4%	-1.1%	20.4%	19.5%	18.4%

Substantial reduction in bias was realized in the simulated data when the optimal set of class weights was used for classification instead of the equal (0.5, 0.5) class weights (17.4 percent). Very little improvement was observed when the true proportions were used for class weights (-1.1 percent). When equal class weights, true proportion, and optimal class weights (i.e., the optimal class weights computed using the simulated data) were used to classify Finney County imagery data, the

D. Conclusions

A procedure for selecting class weights for the maximum likelihood classifier which minimizes the bias in the proportion estimate obtained by sample

counting has been presented. An experiment was run using simulated data and imagery data from Finney County, Kansas. The bias in proportion estimates from classification was compared when using (1) equal class weights, (2) the true proportions for class weights, and (3) optimal class weights. Use of optimal class weights and use of true proportions were found to be superior to the use of equal class weights for this example. Little difference was noted in proportion bias between the use of optimal class weights and the use of true proportions for class weights. It was noted that the classes were well separated in this example, which might explain the lack of differences in the results.

V. REFERENCE

1. Johnson, Ivan L., Jr.: The Davidon-Fletcher-Powell Penalty Function Method: A Generalized Iterative Technique for Solving Parameter Optimization Problems. JSC Internal Note 75-FM-34, May 1975.