Neural Network Approaches Versus Statistical Methods In Classification Of Multisource Remote Sensing Data

Conference Paper *in* IEEE Transactions on Geoscience and Remote Sensing · August 1989

DOI: 10.1109/IGARSS.1989.578748 · Source: IEEE Xplore

3 authors, including:

Jon Atli Benediktsson
University of Iceland
490 PUBLICATIONS **20,087** CITATIONS

SEE PROFILE

Okan K Ersoy
Purdue University
76 PUBLICATIONS **1,469** CITATIONS

SEE PROFILE

Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data

JON A. BENEDIKTSSON, STUDENT MEMBER, IEEE, PHILIP H. SWAIN, SENIOR MEMBER, IEEE, AND OKAN K. ERSOY, MEMBER, IEEE

Abstract—**Neural network learning procedures and statistical classification methods are applied and compared empirically in classification of multisource remote sensing and geographic data. Statistical multisource classification by means of a method based on Bayesian classification theory is also investigated and modified. The modifications permit control of the influence of the data sources involved in the classification process. Reliability measures are introduced to rank the quality of the data sources. The data sources are then weighted according to these rankings in the statistical multisource classification. Four data sources are used in experiments: Landsat MSS data and three forms of topographic data (elevation, slope, and aspect). Experimental results show that the two different approaches have unique advantages and disadvantages in this classification application.**

I. INTRODUCTION

COMPUTERIZED information extraction from remotely sensed imagery has been applied successfully over the last two decades. The data used in the processing have mostly been multispectral data, and the statistical pattern recognition methods (multivariate classification) are now widely known. Within the last decade, advances in space and computer technologies have made it possible to amass large amounts of data about the Earth and its environment. The data are more and more typically not only spectral data but also include, for example, forest maps, ground cover maps, radar data, and topographic information such as elevation and slope data. There may therefore be many kinds of data available from different sources regarding the same scene. These are collectively called multisource data.

We are interested in using all these data to extract more information and get higher accuracy in classification. However, the conventional multivariate classification methods cannot be used satisfactorily in processing multisource data. This is due to several reasons. One is that the multisource data cannot be modeled by a convenient multivariate statistical model since the data are multitype. They can, for example, be spectral data, elevation ranges, and even nonnumerical data such as ground cover classes or soil types. The data are also not necessarily in common units, and therefore scaling problems may arise. Another problem with statistical classification methods is that the data sources may not be equally reliable. This means that the data sources need to be weighted according to their reliability, but most statistical classification methods do not have such a mechanism. This all implies that methods other than the conventional multivariate classification must be used to classify multisource data.

Various heuristic and problem-specific methods have been proposed to classify multisource data. However, we are interested in developing more general methods which can be applied to classify any type of data. In this respect, two approaches will be considered: a statistical approach (parametric) and a neural network approach (distribution-free).

In the statistical case, our attention is focused on statistical multisource analysis by means of a method based on Bayesian classification theory which was proposed by Swain *et al.* [1], [2]. This method will be investigated and extended to take into account the relative reliabilities of the sources of data involved in the classification. This requires a way to characterize and quantify the reliability of a data source, which becomes important when we look at the combination of information. We will investigate methods to determine the reliabilities and to translate them into weights to be used in the classification process.

Recently, there has been a great resurgence of research in neural networks. New and improved neural network models have been proposed, models which can be successfully trained to classify complex data. The generalized delta rule is an example of such a method. In this paper, neural networks for classification of multisource remotely sensed and other geographic data are investigated and compared to the statistical approaches. In the remote sensing community, the question of how well the neural network models perform as classifiers, compared to statistical classification methods, is of considerable interest. Neural network models have as an advantage over the statistical methods that they are distribution-free and no prior knowledge is needed about the statistical distributions of the classes in the data sources in order to apply

Manuscript received October 8, 1989; revised March 1, 1990. This work was supported in part by the National Aeronautics and Space Administration under contract NAGW-925.

The authors are with the School of Electrical Engineering and the Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907.

IEEE Log Number 9036095.

these methods for classification. The neural network methods also take care of determining how much weight each data source should have in the classification. A set of weights describes the neural network, and these weights are computed in an iterative training procedure. On the other hand, neural network models can be very complex computationally, need a lot of training samples to be applied successfully, and their iterative training procedures usually are slow to converge. Also, neural network models have more difficulty than do statistical methods in classifying patterns which are not identical to one or more of the training patterns. The performance of the neural network models in classification is therefore more dependent on having representative training samples, whereas the statistical approaches need to have an appropriate model of each class.

Given the potential advantages of neural networks over statistical methods for classification of multisource data, the purpose of this paper is to determine empirically how well these methods perform as classifiers. Experimental results of classification using both neural network models and statistical methods will be given, and the approaches will be compared based on these results.

II. Statistical Approaches

A. Previous Work

Several statistical methods have been used in the past to classify multisource data. For instance, topographic data have been combined with remotely sensed data in land-cover analysis. One such approach is to subdivide the data into subsets of the data sources and then analyze each subdivision, as reported in Strahler *et al.* [3]. In this method the data are subdivided in such a way that variation within each subdivision is minimized or eliminated based on some of the subdividing variables. Other examples of similar methods can be found in Franklin *et al.* [4] and Jones *et al.* [5].

A second method is "ambiguity reduction," wherein the data are classified based on one or more of the data sources, the results from the classification are assessed, and other sources are then used in order to resolve the remaining ambiguities. The ambiguity reduction can be achieved by logical sorting methods. Hutchinson has used this method successfully [6]. A method related to ambiguity reduction is the layered classifier (tree classifier) applied by Hoffer *et al.* [7]. This particular approach has the advantage that it treats the data sources separately but has the shortcoming that it is very dependent on the analyst's knowledge of the data. Also, as in ambiguity reduction, different groupings or orderings of the sources produce different results [8].

Still another method is supervised relaxation labeling, derived by Richards *et al.* [9] in order to merge data from multiple sources. This method, like other relaxation methods, tries to develop consistency among a collection of observations by means of an iterative numerical "diffusion" process. So far, this method has not been fully investigated on multiple sources and its iterative nature makes it very expensive computationally.

None of the methods described is a general approach to multisource classification, and all of them depend heavily on the user. They all deal with the various sources of data independently. In contrast, a fourth method is a general approach that does not deal with the data sources independently. This method is the stacked-vector approach, i.e., formation of an extended vector with components from all of the data sources and handling the compound vector in the same manner as data from a single source. This method is the most straightforward and conceptually the simplest of the methods. It works very well if the data sources are similar and the relations between the variables are easily modeled [10]. However, the method is not applicable when the various sources cannot be described by a common model, e.g., the multivariate Gaussian model. Another drawback is that when the multivariate Gaussian model is used, the computational cost grows as the square of the total number of variables, which becomes prohibitive if the number of sources is large.

All the methods discussed up to this point have significant limitations as general approaches for multisource classification. Our goal is to develop a general method which can be used to classify complex data sets containing multispectral, topographic, and other forms of geographic data. One such method is focused on in the following. This is a method of statistical multisource analysis, a probabilistic method based on Bayesian decision theory which was developed by Swain *et al.* [1], [2]. The method of statistical multisource analysis will be modified to include mechanisms to control the influence of the data sources in the classification.

B. Fundamentals of Statistical Multisource Analysis

The method proposed in [1] and [2] extends well-known concepts used for classification of multispectral images involving a single data source. In this method the various data sources are handled independently, and each can be characterized by any appropriate model. The data in each source are first classified into source-specific "data classes." The information from the sources is then aggregated by a global membership function, and the data are classified according to the usual maximum selection rule into a user-specified number of "information classes." If n data sources are used, the global membership function has the following form for the information class ω_j:

$$F_j(X) = \left[p(\omega_j)\right]^{1-n} \prod_{s=1}^{n} p(\omega_j | x_s) \qquad (1)$$

where $X = [x_1, x_2, \cdots, x_n]$ is a pixel, $p(\omega_j)$ is the prior probability of ω_j, and $p(\omega_j | x_s)$ is a source-specific posterior probability.

This statistical multisource analysis approach is an extension of single-source Bayesian classification. However, the method as presented by Swain *et al.* [1], [2] does not provide a satisfactory mechanism to account for vary-

ing degrees of reliability among the sources. In our extension of the method, reliability factors are associated with each source involved in the classification. In the following section, we will develop a modified version of this method by means of which reliability analysis is incorporated in the classification process.

1) The Reliability Approach: Consensus theory [11]–[13] is a well-established research field where procedures for combining single probability distributions to summarize the estimates are studied with the assumption that the data sources are Bayesian. In consensus theory, the problem of weighting the various data sources has been studied theoretically [11], [12]. However, the study has mostly been done for "linear opinion pools" whereas, in contrast, (1) can be considered an "independent opinion pool."

We want, in a similar fashion to the weighting in consensus theory, to associate reliability factors with the sources in the global membership function just discussed; i.e., to express quantitatively our confidence in each source and to use the reliability factors for classification purposes. Our goal is to increase the influence of the "more reliable" sources (the sources we have greater confidence in) on the global membership function and consequently decrease the influence of the "less reliable" sources in order to improve the classification accuracy. One reason for employing reliability factors becomes apparent by looking at (1), in which the global membership function is a product of probabilities related to each source. Each probability has a value in the interval from 0 to 1. If any one of them is near zero, it will carry the value of the membership function close to zero and therefore downgrade drastically the contribution of information from other sources, even though the particular source involved may have little or no reliability.

Thus it is desirable to associate weights (reliability factors) with the sources that will influence their respective contributions to classification. Since the global membership function is a product of probabilities, this weight must be involved in such a way that when the reliability of a source is low, the effect is to discount the influence of that source, and when the reliability of a source is high, the effect is to give the source relatively high influence. One possible choice is to introduce reliability factors in the form of exponents on the factor for each source in the global membership function.

Let us now examine the contribution from a single source in the global membership function (1). If one source is added to n sources, i.e., we have $n + 1$ sources, the global membership function could be written in the following form:

$$F_j(X) = \left[p(\omega_j)\right]^{-n} \prod_{i=1}^{n+1} p(\omega_j|x_i). \tag{2}$$

If (2) is divided by (1), we get the contribution from source $n + 1$ which is $p(\omega_j|x_{n+1})/p(\omega_j)$. This motivates us to rewrite (1) in the following form:

$$F_j(X) = p(\omega_j) \prod_{i=1}^{n} \left\{p(\omega_j|x_i)/p(\omega_j)\right\}. \tag{3}$$

Now, to control the influence of each source, reliability factors α_i are assigned as exponents on the contribution from each source. Therefore (3) with reliability factors is written as

$$F_j(X) = p(\omega_j) \prod_{i=1}^{n} \left\{p(\omega_j|x_i)/p(\omega_j)\right\}^{\alpha_i} \tag{4a}$$

where the α_i's ($i = 1, \cdots, n$) are selected in the interval [0, 1]. Notice that if source i is totally unreliable ($\alpha_i = 0$), it will not have any influence on (4a) because

$$\left\{p(\omega_j|x_i)/p(\omega_j)\right\}^0 = 1$$

regardless of the value of $p(\omega_j|x_i)$. And if source i has the highest reliability ($\alpha_i = 1$), then it will give a full contribution to (4a) because

$$\left\{p(\omega_j|x_i)/p(\omega_j)\right\}^1 = p(\omega_j|x_i)/p(\omega_j).$$

It is also worthwhile to note that this method of putting exponents on the probabilities does not change the decision for a single-source classification because the exponential function p^α is a monotonic function of p. A schematic diagram of the statistical multisource classifier, including the reliability factors, is shown in Fig. 1.

Equation (4a) can also be written in a logarithmic form as

$$\log F_j(X) = \log p(\omega_j) + \sum_{i=1}^{n} \alpha_i \log \left\{p(\omega_j|x_i)/p(\omega_j)\right\} \tag{4b}$$

where the reliability factors are expressed as the coefficients in the sum. These coefficients act as weights in the sum and control the influence of each source on the global membership function. Another way to see this is to look at the sensitivity of the global membership function to changes in one of the probability ratios. This can be expressed as

$$\frac{\delta F_j(X)}{F_j(X)} = \alpha_i \frac{\delta p(\omega_j|x_i)/p(\omega_j)}{p(\omega_j|x_i)/p(\omega_j)} \tag{5}$$

and shows that α_i represents the sensitivity; i.e., the value of α_i controls the influence of source i on the global membership function since a percentage change in the posterior probability leads to the same percentage change in the global membership function, multiplied by α_i.

The problem is to quantify the reliability of the sources and define the reliability factors $\{\alpha_i\}$ based on the reliability of the sources. We think of a source as being reliable if its contribution to the combination of information from various sources is "good"; i.e., the classification accuracy is increased.

The process of determining the reliability factors can be characterized as a two-stage process. First, the relia-

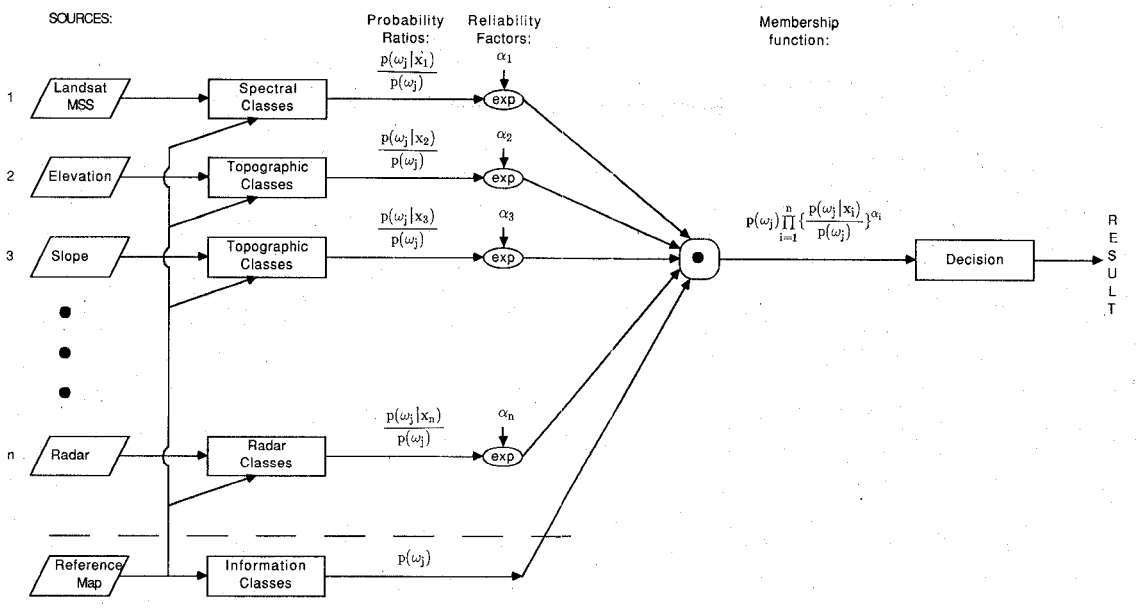


Fig. 1. Schematic diagram of statistical multisource classifier.

bilities of the sources must be quantified by some appropriate "reliability measure," and then the values of the reliability measures must be associated with the reliability factors in the global membership function.

2) Reliability Measures: Using the preceding understanding of a reliable source, three measures are applied to quantify the reliability of a source: weighted average separability, overall classification accuracy, and equivocation. All these measures are related to the classification accuracy of the source.

a) Separability of information classes: We call a source reliable if the separability of the information classes is high for the source. If, on the other hand, the separability of the information classes is low, the source is not reliable. Therefore one possibility for reliability evaluation is to use the average statistical separability of the information classes in each source, e.g., average Bhattacharyya distance [14], average Jeffries–Matusita (JM) distance, average transformed divergence, or any other separability function [15]–[17]. What kind of average is used depends on what we are after in the multisource classification. For instance, if it is desired to improve the overall classification accuracy, the overall average is used. If, however, we are concentrating on specific classes, a weighted average separability of those information classes is used. Using separability with the Gaussian assumption has the disadvantage that it is necessary to compute covariance matrices to estimate the separability. As we shall see later, covariance matrices do not always exist; e.g., for some topographic classes, because the class-specific topographic data can have very little variation. On the other hand, if the Gaussian assumption is not made, it becomes more difficult to compute the separability.

b) Classification accuracy of a data source: Another way to measure reliability of a data source is to use the classification accuracy of the source. In this case a source is considered reliable if the classification accuracy for the source is high, but if the accuracy is low, the source is considered unreliable. This approach is related to the method of using separability measures in that increased separability is consistent with higher accuracy. In contrast to the separability measures, this approach depends only on the classification results and is independent of the classifier or data model used. Since there is no need to estimate covariance matrices to compute the classification accuracy, this approach can always be applied.

c) Equivocation: The motivation for our third reliability measure, equivocation, is to capture how strongly the data classes in the data sources indicate the information classes in the reference map (a map containing "ground truth"). Here we consider a data source very reliable if its data classes strongly indicate the information classes. Conversely, a data source is considered unreliable if there is little or no such indication. The indication can be cast in the form of the conditional probabilities that a specific information class is observed in the reference map, given that there is a specific data class in a classification map from a data source. All these conditional probabilities can be computed by comparing the reference map to a classification map from a data source.

Assuming there are M information classes $\{\omega_1, \cdots, \omega_M\}$ and m data classes $\{d_1, \cdots, d_m\}$, the conditional probabilities can be written as the $m \times M$ correspondence matrix R, where R is

$$R = \begin{bmatrix} p(\omega_1|d_1) & p(\omega_2|d_1) & \cdots & p(\omega_M|d_1) \\ p(\omega_1|d_2) & p(\omega_2|d_2) & \cdots & p(\omega_M|d_2) \\ \vdots & \vdots & \cdots & \vdots \\ p(\omega_1|d_m) & p(\omega_2|d_m) & \cdots & p(\omega_M|d_m) \end{bmatrix}. \quad (6)$$

Reliability can now be characterized in the following way. If a source were perfectly reliable, there would be a unique information class corresponding to each data class. Therefore, ideally, one conditional probability in each row of R would be 1 and all the others would be zero. On the other hand, if a source were very unreliable, there would be no correspondence between the data classes and the information classes; in the worst case, all the numbers in the matrix would be equal.

Now it is necessary to associate a number with the matrix R to quantify the reliability. Using information-theoretic measures [18], the information classes can be thought of as transmitted signals and the data classes as received signals which must be used to estimate the transmitted signals. Using this approach, it can be stated that there is an uncertainty of $\log [1/p(\omega_i|d_j)]$ about the information class ω_i when data class d_j is observed in a data source.

The average loss of information can be calculated when the data class d_j is observed, which is given by

$$H(\omega|d_j) = \sum_i p(\omega_i|d_j) \log \frac{1}{p(\omega_i|d_j)}. \quad (7)$$

The average information loss over all observed data classes d_j is the equivocation of ω with respect to d and is denoted by $H(\omega|d)$:

$$\begin{aligned} H(\omega|d) &= \sum_j p(d_j)\, H(\omega|d_j) \\ &= \sum_i \sum_j p(d_j)\, p(\omega|d_j) \left\{ \log \frac{1}{p(\omega_i|d_j)} \right\} \\ &= \sum_i \sum_j p(d_j)\, p(\omega_i|d_j) \left\{ \log \frac{1}{p(\omega_i|d_j)} \right\}. \end{aligned} \quad (8)$$

$H(\omega|d)$ represents the average uncertainty about an information class over all the data classes. Evidently, $H(\omega|d)$ is the average loss of information per data class, which seems to be a reasonable term to associate with the reliability of a source. Since $H(\omega|d)$ measures uncertainty, the lower its value, the more reliable the source is. Therefore the equivocation is called an uncertainty measure rather than a reliability measure. To be able to transform this uncertainty measure into a reliability factor, it first must be mapped into a reliability measure and then associated with a reliability factor.

3) Association: The values of the reliability (uncertainty) measures must be associated with the reliability factors in order to improve the classification accuracy. It is worthwhile to note that we only want to include sources in the global membership function if the presence of those sources improves the classification accuracy; i.e., we want the classification accuracy to be an increasing function of the number of sources. This is similar to feature selection, but the difference here is that the sources (features) are not only selected but also the contribution of each source to the global membership function is quantified.

Each of the measures discussed in Section II-B2 gives a specific value which should be mapped into a reliability factor on the basis of the strength of the contribution of the source to the classification accuracy. The reliability (or uncertainty) measures take values in some particular interval, and it is necessary to determine a (functional) mapping between the values of the measures and the values of the reliability factors. We have not been able to determine an explicit association function between the values of the reliability and uncertainty measures on one hand and the reliability factors on the other. The measures can easily be used to rank the sources from "best" to "worst," but it is difficult to determine the "optimal" value of the reliability factors. Ranking measures have previously been used in consensus theory [11] for linear opinion pool problems.

In this paper (Section IV), we will not attempt to find the optimal reliability factors. We will simply use the reliability and uncertainty measures to rank the sources. Several different reliability factors will be used in the experiments and we will investigate the relationships between the values of the reliability factors and the overall classification accuracy after combination of the sources.

III. The Neural Network Approach

A neural network is a network of neurons wherein a neuron can be described in the following way: a neuron has many (continuous-valued) input signals x_j, $j = 1, 2, \cdots, N$, which represent the activity at the input or the momentary frequency of neural impulses delivered by another neuron to this input [19]. In the simplest formal model of a neuron, the output value or the frequency of the neuron, o, is often approximated by a function

$$o = K\phi\left(\sum_{j=1}^{N} w_j x_j - \theta \right)$$

where K is a constant and ϕ is a nonlinear function which takes the value 1 for positive arguments and -1 (or 0) for negative arguments. The w_j are called "synaptic efficacies" [19], or weights, and θ is a threshold.

In the neural network approach to pattern recognition, the neural network operates as a black box which receives a set of input vectors $\boldsymbol{x}$ (observed signals) and produces responses o_i from its output units i ($i = 1, \cdots, L$, where L depends on the number of information classes). A general idea followed in neural network theory is that the outputs are either $o_i = 1$, if neuron i is active for the current input vector $\boldsymbol{x}$, or $o_i = -1$ (or 0) if it is inactive. This means the signal values are coded as binary vectors, and for a specific input vector $\boldsymbol{x}$, the outputs give the binary representation of its class number. The process is then to learn the weights through an adaptive (iterative) training procedure. The training procedure is ended when the network has stabilized, i.e., when the weights do not change from one iteration to the next iteration or change less than a threshold amount. Then the data are fed into the network to perform the classification, and the network pro-

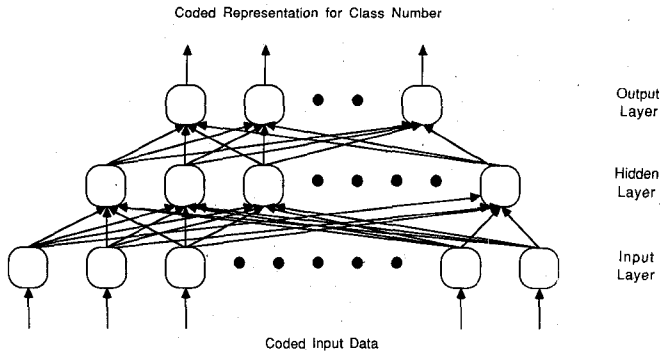


Fig. 2. Schematic diagram of neural network used for classification of image data.

vides at the output the class number of each pixel. A schematic diagram of a three-layer neural network classifier is shown in Fig. 2.

Data representation is very important in application of neural network models. A straightforward coding approach used by most researchers is to code the input and output by a binary coding scheme (0 = 00, 1 = 01, 2 = 11, etc.). However, in some respects for our application, it is more appropriate to use the Gray code representation [20] of the input data. The Gray code representation can be derived from the binary code representation in the following manner: If $b_1, b_2, \cdots b_n$ is a code word in an n-digit binary code, the corresponding Gray code word $g_1, g_2, \cdots g_n$ is obtained by the rule

$$g_1 = b_1$$

$$g_k = b_k \oplus b_{k-1}, \qquad k \geq 2$$

where $\oplus$ is modulo-two addition [20]. The reason that the Gray code representation is more appropriate than the binary code in our application is that adjacent integers in the Gray code differ only by one digit. It can be assumed that adjacent data values in the code space are likely to belong to the same information class. When they belong to the same class, the use of the Gray code leads to a smaller number of weight changes, since for values from a given class, most of the input digits are identical.

Representation at the output of the neural network is also important. If binary coding is used at the output, the number of output neurons can be reduced to $\lceil \log_2 M \rceil$, where M is the number of information classes. However, it is better to use more output units than the minimum $\lceil \log_2 M \rceil$ in order to make the neural network more accurate in classification. Even though adding more output units makes the network larger and therefore computationally more complex, it can also lead to fewer learning cycles, since the Hamming distance of the output representations of different classes can be larger. One such coding mechanism is temperature coding, in which the representation for n has 1 in its first n digits and -1 in the rest (e.g., 4 = 1 1 1 1 -1 -1 -1).

A. Previous Work

Recently, some researchers have applied neural network classifiers to remote sensing data. McClelland *et al.* [21] use a three-layer back propagation algorithm to classify Landsat TM (thematic mapper) data. Decatur [22] uses three-layer back propagation to classify SAR (Synthetic Aperture Radar) data and compares his results to the results of Bayesian classification. Ersoy *et al.* [23] have developed a hierarchical neural network (HNN), which they have applied to classification of aircraft multispectral scanner data. All these researchers report promising performance by neural networks. However, our classification problem is more difficult and our motivation is different. The main reason we are applying neural network methods to the classification of multisource remote sensing data is that these methods are distribution-free. Since multisource data are, in general, of multiple types; the data in each source can have different statistical distributions. By using neural network approaches, we do not have the problem of explicitly modeling the data in each source. Also, the neural network approaches avoid the problem in statistical multisource analysis of specifying how much influence each data source should have on the classification.

Two neural network approaches have been implemented and applied in our research: the delta rule and the generalized delta rule.

B. The Delta Rule

The delta rule, developed by Widrow and Hoff [24] in the early 1960's, is a supervised training approach where error correction is done with a least-mean-squares algorithm (LMS) [25]. The neural network has two layers: input and output layers. The delta rule for updating weights on the kth presentation of an input pattern can be written as

$$\boldsymbol{W}(k) = \boldsymbol{W}(k-1) + \eta\left[\boldsymbol{t}(k) - \boldsymbol{W}(k-1)\,\boldsymbol{x}(k)\right]\boldsymbol{x}^T(k)$$

where $\boldsymbol{x}(k)$ is the input pattern, $\boldsymbol{t}(k)$ is the desired output, $\boldsymbol{W}(k)$ is the state of the weight matrix describing the network after k presentations, and η is a learning rate. Since the magnitudes of the weights change in proportion to η, the optimum learning rate is the one which has the largest value that does not lead to oscillation. A possible choice is $\eta = C/n\text{cycle}$, where C is a constant and $n$cycle the number of the learning cycle. That particular choice of η forces the weight matrix $\boldsymbol{W}(k)$ to stabilize after several iterations. The delta rule, which is identical to the mathematical method of stochastic approximation for regression problems, cannot be used to discriminate data that are not linearly separable and fails, for instance, in the learning of an XOR function.

Since this rule is not even guaranteed to discriminate linearly separable data, it is not expected to perform well in classification of multisource data. However, the delta rule has been generalized to include one or more layers of hidden units. The generalization, which is described

next, can be used to discriminate data which are not linearly separable.

C. The Generalized Delta Rule

The generalized delta rule, or the principle of back propagation of errors, was initially proposed by Werbos in 1974 [26] and later independently developed by Parker in 1986 [27], Le Cun in 1986 [28], and Rumelhart *et al.* in 1986 [29], [30]. The application of the generalized delta rule involves two phases. During the first phase the input is presented and propagated forward through the network to compute the output value o_{pj} in presentation p for each unit j; i.e.,

$$o_{pj} = f_j(\text{net}_{pj})$$

where $\text{net}_{pj} = \Sigma_i w_{ji} o_{pi}$, w_{ji} is the weight of the connection from unit i to unit j, and f_j is the semilinear activation function at unit j which is differentiable and nondecreasing. A widely used choice for a semilinear activation function is the sigmoid function, which is used in our experiments:

$$f_j(\text{net}_{pj}) = 1/(1 + e^{-(\text{net}_{pj} + \theta_j)}).$$

It is worth noting that this function reaches one when net_{pj} goes to infinity and zero when net_{pj} goes to minus infinity. To avoid extremely large values of net_{pj}, the target values of the sigmoid function are usually selected as -0.9 and 0.9 (or 0.1 and 0.9).

The second phase involves a backward pass through the network (analogous to the initial forward pass), during which the error signal δ_{pj} is passed to each unit in the network and the appropriate weight changes are made according to

$$\Delta_p w_{ij} = \eta \delta_{pj} o_{pi}. \tag{9}$$

This second, backward pass allows the recursive computation of δ_{pj} [29]. $\Delta_p w_{ij}$ also gives the negative value of the gradient of the error at the outputs of the neurons multiplied by η. We use the norm of (9) as our convergence criterion for the training process in Section IV-B. When the norm of this scaled gradient is small, there have been little or no weight changes done by the neural network and the network has stabilized.

IV. Experimental Results

The statistical and neural network classification methods were used to classify a data set consisting of the following four data sources:

1) Landsat MSS data (four data channels),
2) elevation data (in 10-m contour intervals, one data channel),
3) slope data (0°–90° in 1° increments, one data channel),
4) aspect data (1°–180° in 1° increments, one data channel).

Each channel comprises an image of 135 rows and 131 columns, all of which are co-registered.

TABLE I
Ground Cover Classes

Class #	Information Class
1	water
2	Colorado blue spruce
3	mountane/subalpine meadow
4	aspen
5	ponderosa pine
6	ponderosa pine/Douglas fir
7	Engelmann spruce
8	Douglas fir/white fir
9	Douglas fir/ponderosa pine/aspen
10	Douglas fir/white fir/aspen

The area used for classification is a mountainous area in Colorado. It has ten ground cover classes, which are listed in Table I. One class is water; the others are forest type classes. It is very difficult to distinguish between the forest types using the Landsat MSS data alone since the forest classes show very similar spectral response. With the help of elevation, slope, and aspect data, they can be better distinguished.

Ground reference data were compiled for the area by comparing a cartographic map to a color composite of the Landsat data and also to a line printer output of each Landsat channel. By this method, 2019 ground reference points (11.4% of the area) were selected. Ground reference consisted of two or more homogeneous fields in the imagery for each class. For each class, the largest field was selected as a training field and the other fields were used for testing the classifiers. Overall, 1188 pixels were used for training and 831 pixels for testing the classifiers.

A. Results: Statistical Approaches

Four statistical methods were used in the experiments performed here: (i) Minimum euclidean distance, (ii) maximum likelihood method for Gaussian data, (iii) Mahalanobis distance for Gaussian data [16], and (iv) statistical multisource classification with the modifications discussed in Section II-B1. The first three methods are "simple" stacked-vector approaches which have been used successfully in classification of remotely sensed data from single sources. It is of interest to see how well these methods work in classification of multisource data. However, we do not expect these methods to give a good performance in terms of classification accuracy for the multisource data, because they are based on the assumption that the data from all sources are modeled in the same way, which is in general not true for multisource data.

To classify the data using the stacked vector methods, the data were clustered, with the initial means being selected as the means of the training classes. The clustering algorithm used (ISODATA algorithm [31]) converged in 13 iterations and gave ten very separable clusters. The mean vectors and the covariance matrices of these ten clusters were then used in the classifications with algorithms 1, 2 and 3. The results of the classifications are shown in Tables II-A and II-B.

TABLE II-A
CLASSIFICATION OF TRAINING SAMPLES AFTER CLUSTERING[a]

	Percent Agreement with Reference for Class										
	1	2	3	4	5	6	7	8	9	10	OA
ED	100	0	0	64	0	37	85	0	0	0	58.2
ML	100	0	0	45	0	37	100	0	0	20	60.9
MD	100	0	0	45	0	37	100	0	0	18	60.8
# of pixels	408	88	45	75	105	126	224	32	25	60	1188

[a]The following statistical methods were used: 1) the minimum euclidean distance (ED), 2) the maximum likelihood method (ML), and 3) the minimum Mahalanobis distance (MD).

TABLE II-B
CLASSIFICATION OF TEST SAMPLES AFTER CLUSTERING[a]

	Percent Agreement with Reference for Class										
	1	2	3	4	5	6	7	8	9	10	OA
ED	95	0	0	28	10	63	56	0	52	0	46.6
ML	95	0	0	26	10	63	90	0	48	0	49.2
MD	95	0	0	26	10	63	93	0	56	0	49.7
# of pixels	195	24	42	65	139	188	70	44	25	39	831

[a]The following statistical methods were used: 1) the minimum euclidean distance (ED), 2) the maximum likelihood method (ML), and 3) the minimum Mahalanobis distance (MD).

Even though the data are not truly Gaussian, the maximum likelihood and the Mahalanobis distance algorithms yielded better results than the euclidean distance algorithm. However, the overall accuracy of the test pixels was only about 49.5% for both the maximum likelihood method and the Mahalanobis distance.

To satisfy the underlying assumptions of the statistical multisource algorithm and the global membership function in (4a) and (4b), we need to show that the data sources can be treated independently in the classification. We do that by looking at the class-specific correlations between all seven data channels using the reference data. The correlations between the data sources are in most cases low. For classes 1 and 2, there is no variation in the topographic data sources, and consequently the correlation is undefined. Since the correlations between the sources are low in all defined cases, we can treat the data sources as independent and use the global membership function in (4a) and (4b) as the classifier.

Each source was treated independently in training. The data classes in the Landsat MSS source were modeled by the Gaussian distribution, where the means and covariance matrices were estimated from the training fields. The other data sources have non-Gaussian data classes. For these sources the normalized histograms of the training fields were used to estimate the density functions.

Statistical multisource classification was performed on the data with varying weights (reliability factors) for the data sources. The results of classification of the training fields are shown in Table III and for the test fields in Table IV. The reliability and uncertainty measures introduced in Section II-B2 were used to rank the data sources. Looking at the classification results for the specific data sources (in both Tables III and IV), these results indicate that the Landsat data is the most reliable source, elevation second, aspect third, and the slope source the least reliable. This is the same ranking indicated in Table V by the equivocation measure. (The separability measures with the Gaussian assumption cannot be used here since some of the data classes in the topographic sources have singular covariance matrices). In all the experiments, the Landsat data were given the highest weights while the weights of the other sources were varied.

Looking at the classification of the training samples (Table III), we see that by combining all the sources with equal weights, the overall classification accuracy is improved to 74.2%; i.e., by more than 6% compared to the best accuracy in the single source classification (Landsat MSS: 67.9%). By lowering the weights on the topographic sources to a certain point, the overall accuracy increases to 78.0%. Therefore by changing the weights of the sources, the classification accuracy of the training samples is improved by 3.8%. This "best" result is achieved when the Landsat source has full weight and the other sources have 40% weight each. It is also very nearly achieved when the Landsat source has full weight, the elevation source has 50% weight, the aspect source has 40% weight, and the slope source has 30% weight (77.9% overall accuracy). That weighting controls the influence from the sources according to the ranking of our reliability measures. Using some other weights that rank the sources in the same way as the reliability measures also gives very good results. In general, the results in Table IV show that the overall classification accuracy can be

TABLE III
CLASSFICATION RESULTS OF TRAINING SAMPLES FOR FOUR DATA SOURCES AND COMPOSITES WITH VARIOUS VALUES OF RELIABILITY FACTOR IN STATISTICAL MULTISOURCE ANALYSIS—LANDSAT MSS, ELEVATION, SLOPE, AND ASPECT DATA ARE COMBINED

	Percent Agreement with Reference for Class										
L E S A[a]	1	2	3	4	5	6	7	8	9	10	OA
Landsat MSS	99	48	0	80	9	69	92	0	0	0	67.9
Elevation	100	0	0	23	17	13	98	0	16	20	58.4
Slope	100	0	0	0	5	64	0	0	0	0	41.5
Aspect	100	0	0	44	42	15	59	0	0	0	53.6
1. 1. 1. 1.	100	98	0	35	35	80	100	0	0	0	74.2
1. .5 .5 .5	100	99	0	65	34	76	94	0	0	62	77.6
1. .4 .4 .4	100	100	11	71	33	73	95	0	0	58	78.0
1. .3 .3 .3	100	100	11	75	27	71	96	0	0	42	76.9
1. .2 .2 .2	100	98	11	75	23	71	96	0	0	26	75.5
1. .1 .1 .1	100	96	18	75	15	66	97	38	0	0	74.2
1. .8 .4 .6	100	99	0	64	37	79	93	0	0	60	77.8
1. .8 .1 .2	100	100	11	74	17	76	95	0	0	35	76.0
1. .6 .4 .5	100	99	4	67	34	76	94	0	0	60	77.8
1. .5 .3 .4	100	100	11	73	33	75	95	0	0	49	77.9
1. .4 .2 .3	100	100	11	75	27	73	96	0	4	38	77.0
1. .3 .1 .2	100	99	11	75	18	74	96	0	4	22	75.4
# of pixels	408	88	45	75	105	126	224	32	25	60	1188

[a]L E S A indicates the weights assigned to the Landsat MSS (l), elevation (e), slope (s), and aspect (a) sources.

TABLE IV
CLASSFICATION RESULTS OF TEST SAMPLES FOR FOUR DATA SOURCES AND COMPOSITES WITH VARIOUS VALUES OF THE RELIABILITY FACTOR IN STATISTICAL MULTISOURCE ANALYSIS—LANDSAT MSS, ELEVATION, SLOPE, AND ASPECT DATA ARE COMBINED

	Percent Agreement with Reference for Class										
L E S A[a]	1	2	3	4	5	6	7	8	9	10	OA
Landsat MSS	97	0	0	0	25	79	97	0	0	0	53.1
Elevation	100	0	0	20	2	21	100	0	8	21	40.4
Slope	86	0	0	0	0	5	33	0	0	0	24.3
Aspect	95	0	0	15	1	6	19	0	0	0	26.7
1. 1. 1. 1.	86	0	0	25	35	92	86	0	0	0	56.0
1. .5 .5 .5	86	0	0	48	45	80	97	0	0	0	57.9
1. .4 .4 .4	86	0	0	52	49	76	97	0	0	0	57.9
1. .3 .3 .3	86	0	0	54	51	63	97	0	0	44	57.4
1. .2 .2 .2	97	0	0	0	54	80	97	0	0	31	59.5
1. .1 .1 .1	93	0	0	0	54	76	97	0	0	26	57.3
1. .8 .4 .6	100	0	0	51	38	84	97	0	0	0	60.8
1. .8 .1 .2	91	0	0	60	48	72	97	0	0	0	58.6
1. .6 .4 .5	86	0	0	51	44	81	97	0	0	0	58.0
1. .5 .3 .4	86	0	0	54	48	74	97	0	0	0	57.5
1. .4 .2 .3	97	0	0	57	51	55	97	0	0	41	58.2
1. .3 .1 .2	95	0	0	0	55	80	97	0	0	33	59.3
# of pixels	195	24	42	65	139	188	70	44	25	39	831

[a]L E S A indicates the weights assigned to the Landsat MSS (l), elevation (e), slope (s), and aspect (a) sources.

TABLE V
EQUIVOCATION OF THE DATA SOURCES

Source	Equivocation	Rank
MSS	0.216955	1
Elevation	0.252676	2
Aspect	0.277244	3
Slope	0.289636	4

improved by reducing the weights of some of the data sources. In Table IV it is also seen that if the weights of the data sources are decreased too much, the overall classification accuracy goes down, as could be expected.

If we look at the results in Table IV, we see very similar results to the ones in Table III. Table IV shows the results of the classification of test fields, and therefore the classification accuracy is lower than in Table III. If the sources

TABLE VI-A
CLASSIFICATION OF TRAINING SAMPLES USING NEURAL NETWORK MODELS[a]

	Percent Agreement with Reference for Class										
	1	2	3	4	5	6	7	8	9	10	OA
A	38	63	18	53	0	58	95	0	0	45	48.1
B	100	15	38	81	2	77	96	50	24	15	71.6
C	100	81	62	77	31	81	100	91	32	97	85.8
D	100	99	71	97	46	95	100	94	92	100	93.0
E	100	100	84	100	57	98	100	91	100	98	95.0
# of pixels	408	88	45	75	105	126	224	32	25	60	1188

[a]The following were the neural network models used: A: the delta rule (learning rate: $0.3/n$cycle); B: the generalized delta rule (learning rate: $0.3/n$cycle, hidden units: 32); C: the generalized delta rule (learning rate: 0.3, hidden units: 32, learning cycles: 200); D: the generalized delta rule (gain factor: 0.3, hidden units: 32, learning cycles: 438, binary coded inputs and outputs); E: the generalized delta rule (learning rate: 0.3, hidden units: 32, learning cycles: 209, Gray coded inputs and temperature coded outputs).

TABLE VI-B
CLASSIFICATION OF TEST SAMPLES USING NEURAL NETWORK MODELS[a]

	Percent Agreement with Reference for Class										
	1	2	3	4	5	6	7	8	9	10	OA
A	25	71	4	17	0	32	97	14	0	36	27.3
B	94	17	33	43	2	43	99	0	0	8	46.3
C	93	67	48	25	1	42	99	0	8	64	49.5
D	97	71	31	31	1	46	97	2	4	44	49.8
E	98	71	29	42	11	41	96	0	0	77	52.5
# of pixels	195	24	42	65	139	188	70	44	25	39	831

[a]The following were the neural network models used: A: the delta rule (learning rate: $0.3/n$cycle); B: the generalized delta rule (learning rate: $0.3/n$cycle, hidden units: 32); C: the generalized delta rule (learning rate: 0.3, hidden units: 32, learning cycles: 200); D: the generalized delta rule (gain factor: 0.3, hidden units: 32, learning cycles: 438, binary coded inputs and outputs); E: the generalized delta rule (learning rate: 0.3, hidden units: 32, learning cycles: 209, Gray coded inputs and temperature coded outputs).

all have equal weights, then the overall accuracy is 56.0%, which is 2.9% greater than the classification accuracy of the best single source (Landsat MSS: 53.1%). This is not as much of an increase as in the case of classification of training data. By lowering the weights on the topographic data sources, the overall classification accuracy is improved to 60.8%, which is 4.8% more than with the equal weights. This best result is achieved when the Landsat source has full weight, the elevation source 80% weight, the aspect source 60% weight, and the slope source 40% weight. This particular weighting ranks the sources in the same way as the reliability measures. Overall accuracy of 60.8% is very good for the classification of test fields in this area.

B. Results: Neural Network Models

The two neural network approaches were implemented in experiments to classify the data. The neural networks were trained with binary or Gray coded input vectors. Since five of the seven data channels take values in the range from 0 to 255, each data channel was represented by eight bits and therefore eight input neurons. The total number of input neurons was $7 \cdot 8 = 56$. Since the number of information classes was ten, the number of output neurons was either selected as four (binary representation) or ten (temperature code representation). The training procedures of the neural networks were said to converge if the norm of the gradient of the error at the outputs was less than 0.01. In all the experiments, the initial learning rate was selected as 0.3, a value that did not lead to oscillation in many cases.

1) Experiments with the Delta Rule: The delta rule did not converge when a learning rate of 0.3 was used for all cycles. By using a decaying learning rate, $0.3/n$cycle, the algorithm converged in 68 cycles. However, the results using the delta rule are clearly not acceptable (see "A" in Tables VI-A and VI-B).

2) Experiments with the Generalized Delta Rule: The generalized delta rule was implemented in experiments with three or more layers (input, output, and hidden layers). Having more than one hidden layer did not improve the classification performance of this neural network, so only the results with three layers are discussed here. The number of hidden units was varied in our initial experiments with the three-layer network. Networks with 16, 32, 48, and 64 hidden units were tried but the performance of the network in terms of classification accuracy was not improved by using more than 32 hidden units. Therefore 32 hidden units were used in all our experiments reported here. Using the learning rate $0.3/n$cycle

with binary coded inputs and outputs, the generalized delta rule converged very fast, i.e., in six cycles. The classification results for this experiment are marked "B" in Tables VI-A and VI-B. Although these results were much better than the ones obtained with the delta rule, even better results could be expected when a constant learning rate is used. The reason for this is that the decaying learning rate forces the weight matrix to converge prematurely, as in the aforementioned delta rule, with respect to the convergence criterion used.

By changing the learning rate to 0.3 and using all the same parameters as before, the generalized delta rule converged extremely slowly. After 200 learning cycles, it was not close to converging but gave much better classification results for the training data in terms of accuracy than when the decaying learning rate was used. These results are marked "C" in Tables VI-A and VI-B. The generalized delta rule with the constant learning rate converged in 438 cycles. The classification results for this case are, up to this point, by far the best for the training data (93.0%), as shown in Table VI-A (see "D").

The preceding experiment was repeated with different coding mechanisms. Using Gray-coded inputs and temperature coded outputs, the generalized delta rule converged in 209 cycles, and the results of the classifications are shown in Tables VI-A and VI-B (see "E"). With these representations, the highest overall classification accuracy for the training data is reached (95.0%). Also, the best overall classification accuracy of test data in the neural network experiments is reached here (52.5%). The classification accuracy of the training data is very satisfactory since the data set is, as said before, very hard to classify accurately. However, the generalized delta rule is computationally complex and usually takes very long to converge in training. Learning, on the other hand, needed fewer iterations and gave better results with the Gray-coded inputs and temperature-coded outputs than the results obtained with binary-coded inputs and outputs.

Although the generalized delta rule is superior to the other methods in classification of training data, it does not do nearly as well in classifying the test data. This illustrates how important it is to select representative training samples when training a neural network. However, the training data used here might be questionable since only one training field was selected for each information class. This implies that each information class only had one subclass! The classification results for the training fields show that if representative training samples are available, the generalized delta rule can do well in classification of multisource data. Significantly, arriving at a truly representative set of training samples can be very difficult in practical remote sensing applications.

V. Conclusion

This empirical evaluation of neural networks versus statistical methods for classification of multisource remote sensing and geographic data has revealed some striking differences.

The neural network model employing the generalized delta rule showed great potential as a pattern recognition method for multisource, remotely sensed data. It is superior to the statistical methods used, in terms of classification accuracy of training data. It has the advantage that it is distribution-free, and we therefore do not have to know anything about the statistical distribution of the data. This is an obvious advantage over most statistical methods requiring modeling of the data, which is difficult when there is no knowledge of the distribution functions or when the data are non-Gaussian. It also avoids the problem of determining how much influence a source should have in the classification, which remains a problem for statistical methods; e.g., statistical multisource analysis, as discussed in the preceding.

However, the generalized delta rule is computationally complex. When the sample size is large, the learning time can be very long. Our experiments also show how important the representation of the data is when using a neural network. Using Gray-coded inputs and temperature-coded outputs gave higher accuracies and required fewer learning cycles than using binary-coded inputs and outputs. To perform well, the neural network models have to be trained by representative training samples. If that can be achieved, our results show that a three-layer net can outperform the statistical methods.

The statistical multisource analysis method proposed in [1] and [2] and modified in this paper worked well for combining multispectral and topographic data in the experiments. The combination of four data sources gave significant improvement in overall classification accuracy compared to single source classification. Using different levels of weights for different sources also showed promise in our experiments in terms of increase in overall classification accuracy. By assigning specific weights (reliability factors) to the data sources, overall classification accuracy was improved about 4–5% from the overall classification accuracy for which equal weights were used.

The statistical multisource classification algorithm requires representative training samples but tends not to be as sensitive to their being representative as do the neural network models. The statistical algorithm outperforms the neural networks in classifying test data since it is provided with more prior knowledge in the form of the statistical model(s) for the data. Carefully modeled density functions make the statistical approach less likely than the neural network models to misclassify samples not seen during training. Also, the statistical multisource algorithm does not require computationally expensive iterative training as do the neural network models. On the other hand, it requires significantly more insight and effort on the part of the analyst.

Two of the suggested reliability measures were employed as ranking criteria for the data sources in statistical multisource analysis. These worked well, but the problem remains: how to find the optimal weights for the data sources? A goal of current research is to frame this problem as an optimization problem and solve it by linear or

nonlinear programming using the training samples. We are also investigating other methods for modeling non-Gaussian data. The histogram approach applied in this paper is very simple and straightforward. However, more complicated approaches might improve upon it in terms of better modeling. The performance of the statistical multisource algorithm is very dependent on careful modeling of the data sources. When the weighting and modeling are done more properly, we can expect the statistical multisource algorithm to perform even better. The other statistical classification algorithms, the stacked vector approaches (euclidean distance, Mahalanobis distance, and the maximum likelihood algorithm for Gaussian data), did not show good performance. The statistical multisource algorithm outperformed them easily, but that was as expected.

The main advantage statistical classification algorithms have over the neural network models is in general that if we know the distribution functions of the information classes, these methods can work very well. In many cases—as, for instance, in multisource classification—we do not always know the distribution functions. Therefore multilayer neural network models can be more appropriate; in particular, if the training process converges in a reasonable amount of time. Faster learning methods for neural networks are a subject of current research. We are also continuing to evaluate the performance of both neural network models and statistical methods using additional data sets. The goal of our research is eventually to combine the two approaches effectively.

Acknowledgment

The Colorado data set was originally acquired, preprocessed, and loaned to the authors by Dr. R. Hoffer, who is now at Colorado State University. They gratefully thank him for giving them access to the data.

References

[1] P. H. Swain, J. A. Richards, and T. Lee, "Multisource data analysis in remote sensing and geographic information processing," in *Proc. 11th Int. Symp. on Machine Processing of Remotely Sensed Data 1985* (West Lafayette, IN), June 1985, pp. 211–217.

[2] T. Lee, J. A. Richards, and P. H. Swain, "Probabilistic and evidential approaches for multisource data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-25, pp. 283–293, May 1987.

[3] A. H. Strahler and N. A. Bryant, "Improving forest cover classification accuracy from Landsat by incorporating topographic information," in *Proc. Twelfth Int. Symp. on Remote Sensing of the Environment* (Ann Arbor, MI), Apr. 1978, pp. 927–942.

[4] J. Franklin, T. L. Logan, C. E. Woodcock, and A. H. Strahler, "Coniferous forest classification and inventory using Landsat and digital terrain data," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-25, pp. 139–149, Jan. 1986.

[5] A. R. Jones, J. J. Settle, and B. K. Wyatt, "Use of digital terrain data in the interpretation of SPOT-1 HRV multispectral imagery," *Int. J. Remote Sens.*, vol. 9, no. 4, pp. 669–682, 1988.

[6] C. F. Hutchinson, "Techniques for combining Landsat and ancillary data for digital classification improvement," *Photogrammetr. Eng. Remote Sens.*, vol. 48, no. 1, pp. 123–130, 1982.

[7] R. M. Hoffer, M. D. Fleming, L. A. Bartolucci, S. M. Davis, and R. F. Nelson, "Digital processing of Landsat MSS and topographic data to improve capabilities for computerized mapping of forest cover types," LARS Tech. Rep. 011579, 1979, Lab. for Appl. of Remote Sensing in cooperation with Dept. of Forestry and Natural Resources, Purdue Univ., W. Lafayette, IN.

[8] H. Kim and P. H. Swain, "Multisource data analysis in remote sensing and geographic information systems based on Shafer's theory of evidence," in *Proc. IGARSS '89, 12th Can. Symp. on Remote Sensing*, vol. 2, 1989, pp. 829–832.

[9] J. A. Richards, D. A. Landgrebe, and P. H. Swain, "A means for utilizing ancillary information in multispectral classification," *Remote Sens. Environ.*, vol. 12, pp. 463–477, 1982.

[10] R. M. Hoffer and staff, "Computer-aided analysis of Skylab multispectral scanner data in mountainous terrain for land use, forestry, water resources and geological applications," LARS Information Note 121275, 1975, Lab. for Appl. of Remote Sensing, Purdue Univ., W. Lafayette, IN.

[11] R. L. Winkler, "The consensus of subjective probability distributions," *Manag. Sci.*, vol. 15, no. 2, pp. B-61–B-75, Oct. 1968.

[12] K. J. McConway, "The combination of experts' opinions in probability assessment: Some theoretical considerations," Ph.D. thesis, University College, London, 1980.

[13] S. French, "Group consensus probability distributions: A critical survey," in *Bayesian Statistics 2*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith, Eds. New York: North Holland, 1985.

[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.

[15] P. H. Swain, "Fundamentals of pattern recognition in remote sensing," in *Remote Sensing—The Quantitative Approach*, P. H. Swain and S. Davis, Eds. New York: McGraw-Hill, 1978.

[16] J. A. Richards, *Remote Sensing Digital Image Analysis—An Introduction*. Berlin: Springer-Verlag, 1986.

[17] S. J. Whitsitt and D. A. Landgrebe, *Error Estimation and Separability Measures in Feature Selection for Multiclass Pattern Recognition*, LARS Publ. 082377, Lab. for Appl. of Remote Sensing, Purdue Univ., W. Lafayette, IN, 1977.

[18] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Chicago: Univ. of Illinois Press, 1963.

[19] T. Kohonen, "An introduction to neural computing," *Neur. Networks*, vol. 1, no. 1, pp. 3–16, 1988.

[20] B. P. Lathi, *Modern Digital and Analog Communication Systems*. New York: Holt, Rinehart, and Winston, 1983.

[21] G. E. McClellan, R. N. DeWitt, T. H. Hemmer, L. N. Matheson, and G. O. Moe, "Multispectral image processing with a three-layer backpropagation network," in *Proc. IJCNN '89* (Washington, DC), 1989, vol. 1, pp. 151–153.

[22] S. E. Decatur, "Application of neural networks to terrain classification," in *Proc. IJCNN '89* (Washington, DC), 1989, vol. 1, pp. 283–288.

[23] O. K. Ersoy and D. Hong, "A hierarchical neural network involving nonlinear spectral processing," presented at IJCNN '89, Washington, DC, 1989.

[24] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in *1960 IRE WESCON Convention Record*. New York: IRE, 1960, pp. 96–104.

[25] J. A. Anderson and E. Rosenfeld, Eds. *Neurocomputing*. Cambridge, MA: MIT Press, 1988.

[26] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. thesis, Harvard Univ., Cambridge, MA, 1974.

[27] D. Parker, "Learning logic," Ctr. Computational Research in Economics and Management Sci., MIT, Cambridge, MA, Tech. Rep. TR-87, 1985.

[28] Y. Le Cun, "Learning processes in an asymmetric threshold network," in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Souli, and G. Weisbruch, Eds. Berlin: Springer-Verlag, 1986.

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representation by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Vol. 1, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: MIT Press, 1986, pp. 318–362.

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

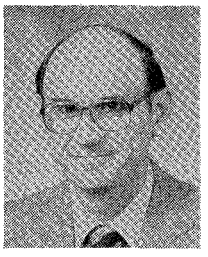
[31] G. H. Ball and D. J. Hall, *A Novel Method of Data Analysis and Pattern Classification*. Menlo Park, CA: Stanford Res. Instit., 1965.

Jon Atli Benediktsson (S'89) received the first degree in electrical engineering from the University of Iceland in 1984. He received the M.S.E.E. degree from Purdue in 1987. He is currently a Ph.D. candidate in the School of Electrical Engineering at Purdue University.

He worked at the Laboratory for Information and Signal Processing at the University of Iceland from 1984 to 1985. Since coming to Purdue in 1985, he has been affiliated with the Laboratory for Applications of Remote Sensing (LARS). His research interests are pattern recognition, image processing, neural networks, and remote sensing.

Mr. Benediktsson is a member of AAAI, INNS, and Tau Beta Pi.

*

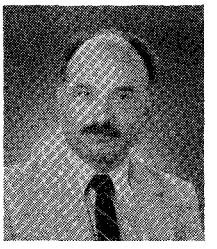


Philip H. Swain (S'66–M'69–SM'81) received the B.S. degree in electrical engineering from Lehigh University in 1963, and the M.S. and Ph.D. degrees from Purdue University in 1964 and 1970, respectively.

He is Professor of Electrical Engineering and Director of Continuing Engineering Education at Purdue University. He has been affiliated with the Laboratory for Applications of Remote Sensing (LARS) at Purdue since its inception in 1966. Much of that time, he served LARS as Program Leader for Data Processing and Analysis Research, responsible for the development of methods and systems for the management and analysis of remote sensing data. He has been employed by the Philco–Ford Corporation and the Burroughs Corporation, and has served as a Consultant to the National Aeronautics and Space Administration (NASA), the Universities Space Research Association, and IBM. During the 1984–1985 academic year, he was an Honorary Visiting Fellow at the University of New South Wales, Sydney, Australia. His research interests include theoretical and applied pattern recognition, methods of artificial intelligence, geographic information systems, and the application of advanced computer architectures to image processing. He is co-editor and contributing author of the textbook *Remote Sensing: The Quantitative Approach* (New York: McGraw–Hill, 1978).

Dr. Swain is a member of the American Society for Engineering Education (ASEE), Phi Beta Kappa, Sigma Xi, and Eta Kappa Nu. As Vice Chairman of the Technical Committee on Remote Sensing (TC7) of the International Association for Pattern Recognition, he was Co-Organizer of a Workshop on Analytical Methods in Remote Sensing for Geographic Information Systems (Paris, France, 1986) and published its proceedings as a special issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

*



Okan K. Ersoy (M'86) received the B.S.E.E. degree from Robert College in 1967, and the M.S., Certificate of Engineering, and the M.S. and Ph.D. degrees from the University of California, Los Angeles (UCLA), in 1968, 1971, and 1972, respectively.

He was a Teaching and Research Assistant in the Department of Electrical Sciences and Engineering, UCLA (1968–1972), Assistant Professor in the Department of Electrical Engineering, Bosphorus University (1972–1973), and Associate Professor in the second semesters at the same university (1976–1980). He joined the Center for Industrial Research, Oslo, Norway, as a Researcher in the Computer Science Division in 1973. He was a Visiting Scientist at UCSD in 1980–1981. He has been with Purdue University, School of Electrical Engineering, West Lafayette, IN, as Associate Professor since August 1985. He was a Fulbright Fellow in 1967–1968. His current interests include digital signal/image processing and understanding, neural computing and information processing, fast VLSI algorithms and architectures, optical computing and signal processing, and holography.