Hybrid Consensus Theoretic Classification

Jon Atli Benediktsson, *Member, IEEE,* Johannes R. Sveinsson, *Member, IEEE,* and Philip H. Swain, *Senior Member, IEEE*

***Abstract*— Hybrid classification methods based on consensus from several data sources are considered. Each data source is at first treated separately and modeled using statistical methods. Then weighting mechanisms are used to control the influence of each data source in the combined classification. The weights are optimized in order to improve the combined classification accuracies. Both linear and nonlinear optimization methods are considered and used in classification of two multisource remote sensing and geographic data sets. A nonlinear method which utilizes a neural network gives excellent experimental results. The hybrid statistical/neural method outperforms all other methods in terms of test accuracies in the experiments.**

I. Introduction

DATA FUSION [1] for classification of remote sensing and geographic data from multiple data sources is a very challenging research problem. Typically, the multisource data now not only include spectral data but also, for example, forest maps, ground cover maps, radar data and topographic information such as elevation and slope. It is desirable to use all of these data to extract more information and achieve higher accuracy in classification.

The major methods proposed in data fusion are based on statistical methods [2]–[4], neural networks [5]–[7], Dempster-Shafer theory [4], [8], and fuzzy logic [9]. Here we will concentrate on the statistical and neural network methods and the combination of those approaches.

Several statistical methods have been used in the past to classify multisource remote sensing data. Conceptually, the simplest method is the stacked-vector approach in which a compound vector with components from all of the data sources is found and processed using the conventional statistical classification techniques in the same manner as data from a single source. This method is very straightforward and works very well if the data sources are similar and the relations among 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 and becomes prohibitive if the total number of variables is large.

Manuscript received September 13, 1996; revised March 3, 1997. This work was supported in part by the Icelandic Research Council and the Research Fund of the University of Iceland.

J. A. Benediktsson and J. R. Sveinsson are with the Engineering Research Institute, University of Iceland, 107 Reykjavik, Iceland (e-mail: benedikt@verk.hi.is).

P. H. Swain is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA.

Publisher Item Identifier S 0196-2892(97)04474-4.

Another approach is to divide the data into subsets of sources and then analyzes each subset [11]–[13]. In this approach, the data are subdivided in such a way that variation within each subdivision is minimized or eliminated based on some of the subdividing variables.

Still another method is ambiguity reduction where the data are classified based on one or more of the data sources, the results of the classification are assessed, and other sources are then used to resolve the remaining ambiguities. The ambiguity reduction can be achieved by logical sorting methods [14]. Another method close to ambiguity reduction is the layered classifier (tree classifier) applied in [15]. This approach has the advantage that it treats the data sources separately but has the shortcomings that design of the classifier is very dependent on the analyst's expertise and knowledge of the data. Also, as in the ambiguity reduction method, different groupings or orderings of the sources may produce different results [16]. In [17] a knowledge-based method for multisource classification has been proposed but such methods can be very heuristic. Woodcock *et al*. [18] have also proposed a method using mapping systems that resemble knowledge based systems.

Supervised relaxation labeling has been used to merge data from multiple sources [19]. This method, like other relaxation methods, tries to develop consistency among a collection of observations by means of an iterative numerical diffusion process.

It is important to develop general methods which can be used to classify complex data sets containing multispectral, topographic and other forms of geographic data. Solberg *et al.* [20] has used a Markov Random Field model for multisource classification of remotely sensed data. Related to this problem, Solberg *et al.* [3] has proposed statistical techniques for multitemporal data by allowing changes in the identities of the classes and by considering the class-dependent likelihood of changes. Benediktsson and Swain [2] have proposed the use of consensus theory which is a general approach that can be used for any type of data. In their approach the contribution of each expert needs to be weighted. No optimal way of doing that has been proposed.

Neural networks have shown promise in data fusion of multisource data. Benediktsson *et al*. [5] used backpropagation successfully and compared their results to statistical methods. Wan and Fraser [6] have proposed a method which can be used for data fusion, multitemporal classification, and contextual classification. Carpenter *et al*. [7] have proposed the use of ART neural networks in classification of multisource data.

The advantage of neural network methods is that no prior statistical information is needed about the input data. This is very important, since the whole multiple source data set

is usually very difficult to model by statistical methods [5]. However, when a sufficiently accurate statistical model can be determined, statistical methods should outperform neural networks in terms of classification accuracies. In multisource classification it is relatively easy to model each source individually. In this paper, statistical methods are used to model the individual sources. Then, some type of scheme is used to combine the source specific classifications. This is done by the combination of statistical consensus theory and neural networks. The goal of statistical consensus theory is to get a consensus among experts. Here, an individual expert bases his opinion on a specific data source. The neural networks are then used to obtain the consensus among the experts, i.e., the neural networks optimize the consensus with respect to classification accuracy.

The paper is organized as follows. First, consensus theory is reviewed. Then, weight selection schemes in consensus theory are discussed in Section III. Experimental results for two multisource and geographic data sets are given in Section IV. Finally, conclusions are presented in Section V.

II. Consensus Theory

Consensus theory [2], [21]–[25] is a well-established research field involving procedures with the goal of combining single probability distributions to summarize estimates from multiple experts (data sources) with the assumption that the experts make decisions based on Bayesian decision theory. Consensus theory is closely related to the method of stacked generalization [26] where outputs of experts are combined in a weighted sum with weights which are based on the individual performance of the experts. In most consensus theoretic methods each data source is at first considered separately. For a given source an appropriate training procedure can be used to model the data by a number of source-specific densities that will characterize that source [2]. The data types are assumed to be very general. The source-specific classes or clusters are therefore referred to as data classes, since they are defined from relationships in a particular data space. In general there may not be a simple one-to-one relation between the user-desired information classes and the set of data classes available since the information classes are not necessarily a property of the data. In consensus theory, the information from the data sources is aggregated by a global membership function, and the data are classified according to the usual maximum selection rule into the information classes. The combination formula obtained is called a consensus rule.

Consensus theory can be justified by the fact that a group decision is better in terms of mean square error than a decision from a single expert (data source). To show this, let us define an indicator function

$$I_{\omega_j} = \begin{cases} 1, & \text{if } \omega_j \text{ occur} \\ 0, & \text{if } \omega_j \text{ does not occur} \end{cases}$$

where ω_j is an information class. Now it is needed to find an estimate, r, of the "best" probability that minimizes the mean square error (summed over all Z's)

$$\varepsilon^2_{\omega_j}(r) = \sum_Z (r - I_{\omega_j})^2 p(Z)$$

where $Z = [z_1, \cdots, z_n]$ is a compound vector consisting of observations from all the data sources, n is the number of data sources, z_i $(i = 1, \cdots, n)$ is an observation from a single data source (z_i can be a vector if the corresponding data source makes a multidimensional observation), and $p(Z)$ is the probability of Z. Differentiating $\varepsilon^2_{\omega_j}(r)$ with respect to r and setting the result equal to zero gives

$$\sum_Z 2(r - I_{\omega_j}) p(Z) = 0.$$

The solution to the above equation is $r = p(\omega_j \mid Z)$ which implies that the group probability $p(\omega_j \mid Z)$ is optimal for classification in the mean square sense.

Although the above justification for consensus theory exits, it is important to note that this justification does not place any constraints on the choice of the consensus rule (global membership function), and consequently several consensus rules have been proposed in the literature. Probably the most commonly used consensus rule is the linear opinion pool (LOP) (see Fig. 1) which has the following (group probability) form for the information class ω_j if n data sources are used:

$$C_j(Z) = \sum_{i=1}^{n} \lambda_i p(\omega_j \mid z_i) \tag{1}$$

where $p(\omega_j \mid z_i)$ is a source-specific posterior probability and λ_i's $(i = 1, \cdots, n)$ are source-specific weights which control the relative influence of the data sources. The weights are associated with the sources in $C_j(Z)$ (the global membership function for the LOP) to express quantitatively the goodness of each source [22].

The linear opinion pool has a number of appealing properties. For example, it is simple, yields a probability distribution, and the weight λ_i reflects in some way the relative expertise of the ith expert. Also, if the data sources have absolutely continuous probability distributions, the linear opinion pool gives an absolutely continuous distribution. In using the linear opinion pool, it is assumed that all of the experts observe the input vector Z. Therefore, (1) is simply a weighted average of the probability distributions from all the experts and the result is a combined probability distribution.

The linear opinion pool, though simple, has several weaknesses [21]; e.g., it shows dictatorship when Bayes' theorem is applied, i.e., only one data source will dominate in making a decision. It is also not externally Bayesian (does not obey Bayes' rule). The reason it is not externally Bayesian is that the linear opinion pool is not derived from the joint probabilities using Bayes' rule. Another consensus rule, the logarithmic opinion pool (LOGP), has been proposed to overcome some of the problems with the linear opinion pool. The logarithmic opinion pool can be described by

$$L_j(Z) = \prod_{i=1}^{n} p(\omega_j \mid z_i)^{\lambda_i} \tag{2}$$

or

$$\log(L_j(Z)) = \sum_{i=1}^{n} \lambda_i \log(p(\omega_j \mid z_i)) \tag{3}$$

where $\lambda_1, \cdots, \lambda_n$ are weights which should reflect the goodness of the data sources and $L_j(Z)$ is the global membership function for the LOGP.

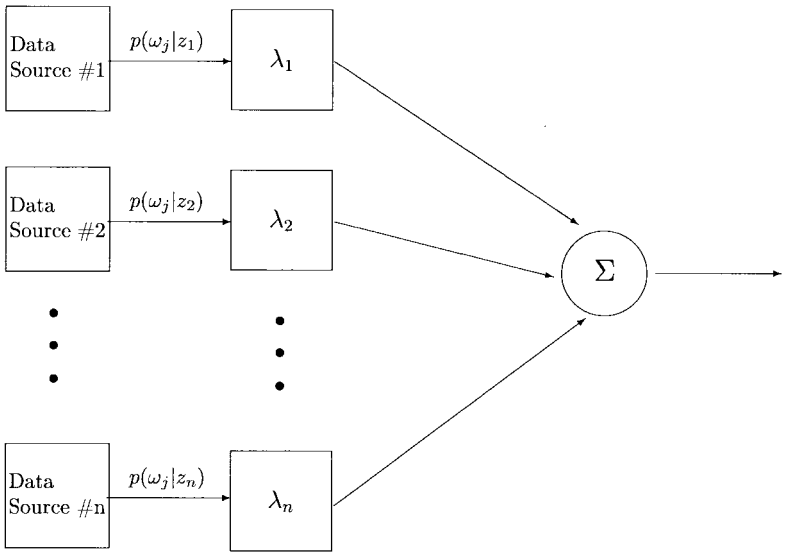

Fig. 1. Schematic diagram of a linear opinion pool.

The logarithmic opinion pool differs from the linear opinion pool in that it is unimodal and less dispersed. Also, the logarithmic opinion pool treats the data sources independently. Zeros in it are vetos; i.e., if any expert assigns $p(\omega_j \mid z_i) = 0$, then $L_j(Z) = 0$. This dramatic behavior is a drawback if the density functions are not carefully estimated. The logarithmic opinion pool is externally Bayesian, but it is computationally more complicated than the linear opinion pool.

III. Weight-Selection Schemes

The previous section focused on consensus rules, but the weight selection schemes for these rules were not addressed. The weight-selection schemes in consensus theory should reflect the goodness of the separate input data sources, i.e., more influence in the decision making should be given to data sources that contribute to higher accuracy. There are at least two potential weight selection schemes. The first scheme is a classical scheme, i.e., to select the weights such that they weight the individual data sources but not the classes within the sources, e.g., use reliability measures which rank the data sources according to their goodness (heuristic weight-selection scheme). These reliability measures might be, e.g., source-specific classification accuracy of training data, overall separability or equivocation [2]. Another version of the classical scheme is the equal weighting method, i.e., all the sources are weighted equally.

The second scheme is to choose the weights such that they not only weight the individual data sources but also the classes within the sources. This scheme consists of defining a function

$$Y = f(X, \Lambda)$$

where X contains source-specific posteriori discriminative information and Λ corresponds to the source-specific weights in (1) and (3).

If f is linear, the combined output response, Y, can be written in matrix form as

$$Y = X\Lambda$$

where X is a matrix containing the discriminative outputs of the consensus theoretic classifier and Λ contains all the weights (see Fig. 2). Assuming that X has full column rank, the above equation can be solved for Λ using the pseudo-inverse of X or a simple linear delta rule. In order to find the optimal weights for the case f is linear, we define

$$X = [X_1\ X_2\ \cdots\ X_n],$$
$$\Lambda = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \\ \vdots \\ \Lambda_n \end{bmatrix}$$

where $X_i\ i = 1, \cdots, n$ are $N \times M$ matrices (N is the number of training samples, and M is the number of information classes). Each row of X_i represents an output vector for the ith data source. The matrices $\Lambda_i,\ i = 1, \ldots, n$ are $M \times M$, representing the weights for the ith data source. If $Y = D$ is the desired output for the whole classification problem, we have

$$X\Lambda = D$$

where Λ is an unknown matrix, and its least square estimate Λ_{lopt} is sought by minimizing the squared error:

$$\Lambda_{\text{lopt}} = \arg\min_{\Lambda} \|X\Lambda - D\|^2. \tag{4}$$

The solution for Λ_{lopt} in (4) can usually be obtained by using the pseudo-inverse of X, i.e.,

$$\Lambda_{\text{lopt}} = (X^T X)^{-1} X^T D \tag{5}$$

where X^T is the transpose of X, and $(X^T X)^{-1} X^T$ is the pseudo-inverse of X which exists if $X^T X$ is nonsingular. In the case that X is not of full column rank, this solution becomes ill-conditioned. In that case one can use dummy augmentation to make X a full column rank matrix in a higher dimensional space and then solve the problem. There

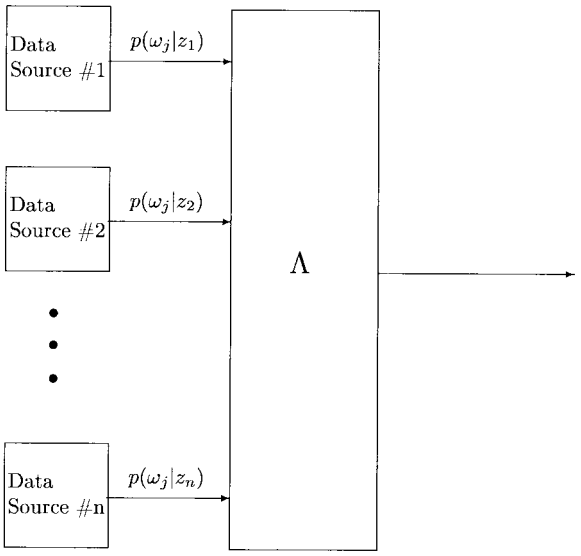

Fig. 2. Schematic diagram of a linearly optimized LOP.

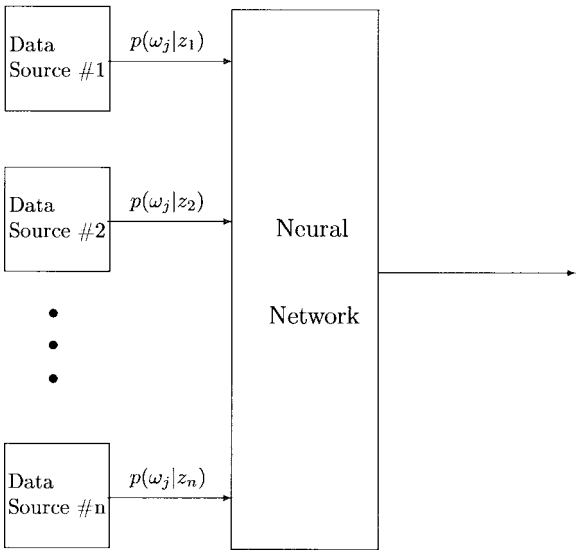

Fig. 3. Schematic diagram of a LOP optimized by a neural network.

are at least two other suboptimal methods for solving this optimization problem.

The first method is to use sequential formulas to compute the optimal Λ [27]. Let the ith row vector of the matrix X be x_i^T and the ith row of the matrix D be d_i^T; then Λ can be calculated iteratively using the formula

$$\Lambda^{(i+1)} = \Lambda^{(i)} + P^{(i+1)} x_{i+1} \left(d_{i+1}^T - x_{i+1}^T \Lambda^{(i)} \right)$$

$$P^{(i+1)} = P^{(i)} - \frac{P^{(i)} x_{i+1} x_{i+1}^T P^{(i)}}{1 + x_{i+1}^T P^{(i)} x_{i+1}} \quad i = 0,\ 1, \cdots, N$$

where $\Lambda^{(N)}$ is the least squares estimate of Λ_{lopt}. The initial conditions to the sequential formula are $\Lambda^{(0)} = 0$ and $P^{(0)} = \beta I$, where β is a positive large number.

The second method for solving the least squares error problem is to choose unitary Λ which minimizes $||D - X\Lambda||^2$ [28]. We compute

$$||D - X\Lambda||^2 = ||D||^2 - 2\langle D, X\Lambda \rangle + ||X||^2$$

where $\langle D, X\Lambda \rangle = \text{tr}(D\Lambda^T X^T)$ and tr returns the trace of its argument matrix. If

$$X^T D = V \Sigma U^T$$

is a singular value decomposition (SVD) of $X^T D$ then

$$\begin{aligned} \text{tr}(D\Lambda^T X^T) &= \text{tr}(X^T D \Lambda^T) \\ &= \text{tr}(V \Sigma U^T \Lambda^T) \\ &= \text{tr}(\Sigma U^T \Lambda^T V) \\ &= \sum_{i=1}^{M} \sigma_i (X^T D) t_{ii} \end{aligned}$$

where $T = [t_{ij}] = U^T \Lambda^T V$ is a unitary matrix and σ_i is the ith singular value of its argument matrix. This sum is maximized when all $t_{ii} = 1$, i.e., when $\Lambda_{\text{lopt}} = VU^T$.

In the case when f is nonlinear, a neural network can be used to obtain a mean square estimate of the function. Here, the consensus theoretic classifiers with equal weights can be considered to preprocess the data for the neural networks. Then, the neural network learns the mapping from the source-specific posteriori probabilities to the information classes. Therefore, the neural network is used to optimize the classification capability of the consensus theoretic classifiers (see Fig. 3). If $Y = D$ is the desired output for the whole classification problem, the process can be described by the equation

$$\Lambda_{n\text{lopt}} = \arg\min_{\Lambda} ||D - f(X, \Lambda)||^2. \tag{6}$$

The update equation for the weights of the neural network is

$$\Delta\Lambda = \eta ||D - f(X, \Lambda)|| \nabla_\Lambda f$$

where η is a learning rate and ∇_Λ is the gradient with respect to Λ.

TABLE I
TRAINING AND TEST SAMPLES FOR INFORMATION CLASSES IN THE EXPERIMENT ON THE COLORADO DATA SET

Class #	Information Class	Training Size	Test Size
1	Water	301	302
2	Colorado Blue Spruce	56	56
3	Mountane/Subalpine Meadow	43	44
4	Aspen	70	70
5	Ponderosa Pine 1	157	157
6	Ponderosa Pine/Douglas Fir	122	122
7	Engelmann Spruce	147	147
8	Douglas Fir/White Fir	38	38
9	Douglas Fir/Ponderosa Pine/Aspen	25	25
10	Douglas Fir/White Fir/Aspen	49	50
Total		1008	1011

IV. EXPERIMENTAL RESULTS

Experiments were conducted on two multisource remote sensing and geographic data sets with different characteristics. Both data sets were from forest areas but the class-conditional information in one data set (Anderson River data) were much

TABLE II
TRAINING ACCURACIES IN PERCENTAGE FOR THE CLASSIFICATION METHODS APPLIED TO THE COLORADO DATA SET

Method	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Average Accuracy	Overall Accuracy
MED	41.5	98.2	25.6	37.1	37.6	0.0	73.5	0.0	40.0	24.5	37.80	40.28
LOP (equal weights)	100.0	0.0	0.0	92.9	38.9	49.2	100.0	0.0	12.0	100.0	49.29	68.06
LOP (heuristic weights)	100.0	25.0	16.3	91.4	36.3	90.2	99.3	0.0	0.0	100.0	55.84	74.21
LOP (optimal linear weights)	100.0	62.5	25.6	74.3	66.2	79.5	98.6	23.7	40.0	91.8	66.23	80.26
LOP (optimized with CGBP)	100.0	87.9	26.2	81.8	67.8	73.0	100.0	39.5	75.0	94.4	74.55	83.46
LOGP (equal weights)	99.7	96.4	20.9	87.1	60.5	46.7	100.0	44.7	44.0	91.8	69.20	78.97
LOGP (heuristic weights)	99.7	91.1	23.3	95.7	45.2	83.6	100.0	5.3	48.0	100.0	69.18	80.46
LOGP (optimal linear weights)	100.0	67.9	23.3	81.4	58.6	82.8	98.6	18.4	28.0	91.8	65.08	79.66
LOGP (optimized with CGBP)	100.0	80.4	69.8	99.6	78.3	82.8	100.0	80.3	100.0	100.0	89.12	91.39
CGBP (0 hidden neurons)	100.0	71.7	67.4	85.2	67.4	73.2	100.0	34.2	65.3	95.9	76.05	84.16
CGBP (40 hidden neurons)	100.0	96.7	95.7	99.5	90.3	89.5	100.0	87.3	96.7	100.0	95.58	96.26
Number of Samples	301	56	43	70	157	122	147	38	25	49		1008

more difficult to model statistically than for the other (Colorado data). In the experiments it was investigated how the difficulty with statistical modeling affected the weight selection schemes for the different consensus rules. The results of the experiments are discussed below.

A. Experiment 1: Colorado Data Set

Classification was performed on a data set consisting of the following four data sources [5], [10], [15]:

1) Landsat MSS data (four spectral 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 comprised an image of 135 rows and 131 columns, and all channels were spatially co-registered. 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 types. It is very difficult to distinguish among the forest types using the Landsat MSS data alone since the forest classes show very similar spectral response [5], [10], [15]. 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 reference points (11.4% of the area) were selected comprising two or more homogeneous fields in the imagery for each class. Approximately 50% of the reference samples were used for training, and the rest were used to test the classification methods.

Three statistical methods were used to classify the data: the minimum Euclidean Distance (MED) classifier [19], the linear opinion pool (LOP), and the logarithmic opinion pool (LOGP). For the LOP and LOGP, ten data classes (corresponding to the information classes in Table I) were defined in each data source. The multispectral remote sensing data sources were modeled to be Gaussian but the topographic data sources were modeled by Parzen density estimation [30] with Gaussian kernels. In this experiment, some of the class-conditional probability density functions were relatively sharp. Consequently, zero class-conditional probabilities were obtained for some information classes in (1) and (3). Therefore, the veto property in (3) was sure to play a role in the consensus theoretic classifications in the experiment.

Several different weighting schemes were tried for the LOP and LOG. These weighting schemes were: 1) equal weights; 2) heuristic weights based on reliability measures [10]; 3) optimal linear weights; and 4) optimal nonlinear weights based on (6). For the optimal nonlinear weights, two and three layer conjugate-gradient backpropagation (CGBP) neural networks [31] were utilized with different numbers of hidden neurons (0, 15, 25, 35, and 45 hidden neurons). For each implementation, the neural networks were trained six times with different initializations. Then, the average accuracy for these six experiments was computed.

The CGBP algorithm with two and three layers was also trained on the same data with different numbers of hidden neurons (0, 15, 30, and 45 hidden neurons). Each version of the CGBP network was also trained six times with different initializations and the overall average accuracies were computed in each case.

The overall classification accuracies for the different methods are summarized in Tables II and III. In the tables the average result for the best implementation of the neural network-based methods is shown in each case. There the LOP and LOGP were both optimized by the CGBP with 30 hidden neurons. Also, the following heuristic weights were used for the LOP in (1): Landsat MSS: 1.0, elevation data: 0.4, slope data: 0.4, and aspect data: 0.4. The heuristic weights for the LOGP in (3) were: Landsat MSS: 1.0, elevation data: 0.6, slope data: 0.6, and aspect data: 0.6. For the consensus theoretic methods that utilize optimal linear weighting, the weighting was determined by the pseudo-inverse method in Section III.

From Table II (training data) it can bee seen that the CGBP with 40 hidden neurons is the most accurate classification method both in terms of average (over the classes) and overall (over the samples) training accuracies. As expected, the MED classifier did not achieve high accuracies. For the

TABLE III
TEST ACCURACIES IN PERCENTAGE FOR THE CLASSIFICATION METHODS APPLIED TO THE COLORADO DATA SET

Method	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Average Accuracy	Overall Accuracy
MED	40.1	100.0	34.1	30.0	32.5	0.8	69.4	0.0	28.0	20.0	35.49	37.98
LOP (equal weights)	100.0	0.0	0.0	87.1	35.0	48.4	100.0	0.0	0.0	94.0	46.45	66.37
LOP (heuristic weights)	100.0	30.4	18.2	80.0	35.7	88.5	100.0	0.0	0.0	96.0	54.87	73.39
LOP (optimal linear weights)	100.0	80.4	25.0	77.1	66.3	75.4	99.3	15.8	32.0	92.0	66.14	80.22
LOP (optimized with CGBP)	100.0	90.2	39.2	75.3	61.0	74.6	99.3	34.9	58.0	96.5	72.90	82.22
LOGP (equal weights)	99.3	100.0	18.2	85.7	56.7	52.5	99.3	42.1	44.0	92.0	68.98	78.73
LOGP (heuristic weights)	100.0	96.4	18.2	91.4	40.8	87.7	99.3	10.5	24.0	100.0	66.84	79.62
LOGP (optimal linear weights)	100.0	76.8	25.0	75.7	63.7	81.1	99.3	13.2	16.0	92.0	64.28	80.02
LOGP (optimized with CGBP)	99.8	64.3	58.0	73.9	61.5	71.7	98.6	49.3	80.0	94.0	75.12	82.27
CGBP (0 hidden neurons)	100.0	71.4	63.3	72.9	56.7	72.8	98.5	25.0	38.0	87.0	68.56	79.72
CGBP (40 hidden neurons)	99.89	57.1	61.0	67.6	59.3	69.1	97.4	34.6	45.3	78.7	67.01	78.37
Number of Samples	302	56	44	70	157	122	147	38	25	50		1011

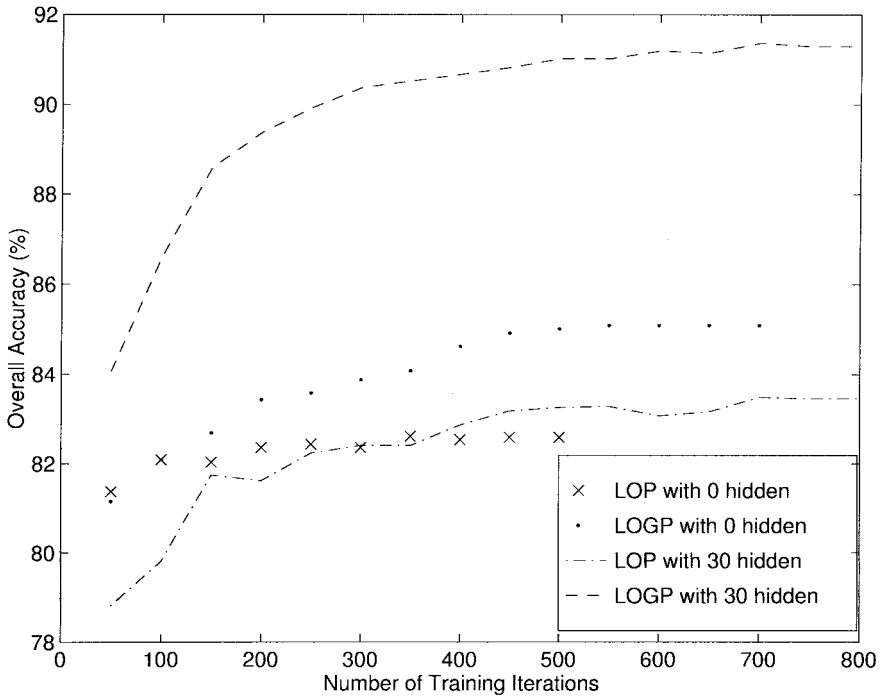

Fig. 4. Colorado data: Training accuracies as a function of the number of iterations for the consensus theoretic methods optimized by the CGBP. Results with and without hidden neurons are displayed for each consensus rule.

consensus theoretic methods, the LOPG optimized with CGBP achieved the highest overall and average training accuracies. The overall training accuracies obtained by the best LOP (also optimized by CGBP) were approximately 8% lower than the ones achieved by the LOGP. Both CGBP optimized versions of the consensus theoretic methods achieved significant improvements over their equally weighted versions. In the case of the LOP optimized by the CGBP, the training accuracy was improved by more than 15% but for the CGBP-optimized-LOGP it increased by over 12%.

In Table III (test data) it can be seen that the CGBP optimization increased the test accuracies of the equally weighted LOP by nearly 16% but the equally weighted LOGP by almost 6%. The CGBP optimized versions of the consensus theoretic classifiers were the most accurate classifiers considered in terms of classification accuracy of test data. The results for these methods were similar in terms of overall test accuracies but the LOGP achieved higher average test accuracies (classification accuracy for class 9 is the major cause for this difference). The LOP and LOGP with CGBP optimization outperformed both versions of the neural network classifiers in terms of overall and and average test accuracies. This indicates the significant result that the discriminative information provided by equally weighted consensus theoretic classifiers are effective preprocessors for neural networks.

In Figs. 4 and 5 the overall accuracies for the CGBP optimized consensus theoretic classifiers are shown with both zero and 30 hidden neurons. There it is seen that the hidden neurons are more important in the neural network optimization

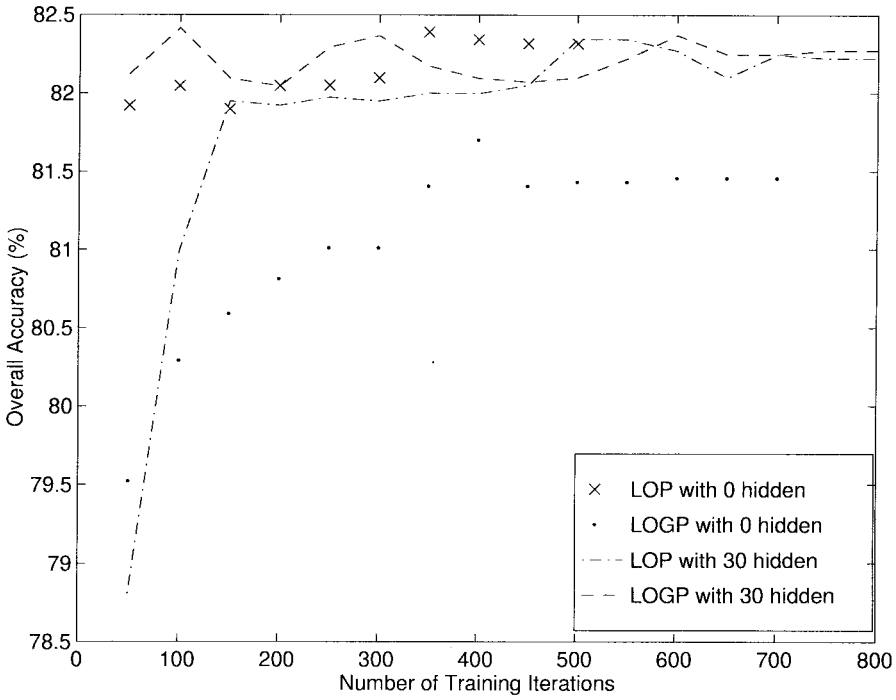

Fig. 5. Colorado data: Test accuracies as a function of the number of iterations for the consensus theoretic methods optimized by the CGBP. Results with and without hidden neurons are displayed for each consensus rule.

of the LOGP than the LOP. This comes as no surprise since the LOGP, in contrast to the LOP, is a nonlinear consensus rule.

B. Experiment 2: Anderson River Data Set

The data used in the second experiment, the Anderson River data set, are a multisource remote sensing and geographic data set made available by the Canada Centre for Remote Sensing (CCRS) [32]. This data set is very difficult to classify. The imagery involves a 2.8 $\times$ 2.8 km forestry site in the Anderson River area of British Columbia, Canada. The area is characterized by rugged topography, with terrain elevations ranging from 330 m to 1100 m above sea level. The forest cover is primarily coniferus, with Douglas fir predominating up to approximately 1050 m elevation, and cedar, hemlock and spruce types predominating at higher elevations [32]. Six data sources were used:

1) Airborne Multispectral Scanner (AMSS) with 11 spectral data channels (ten channels from 380 to 1100 nm and one channel from 8 to 14 μm).
2) Steep Mode Synthetic Aperture Radar (SAR) with four data channels (X-HH, X-HV, L-HH, L-HV).
3) Shallow Mode SAR with four data channels (X-HH, X-HV, L-HH, L-HV).
4) Elevation data (one data channel, where elevation in meters $= 61.996 + 7.2266 *$ pixel value).
5) Slope data (one data channel, where slope in degrees $=$ pixel value).
6) Aspect data (one data channel, where aspect in degrees $= 2 *$ pixel value).

The AMSS and SAR data were detected during the week of July 25 to 31, 1978. Each channel comprises an image of 256 lines and 256 columns. All of the images are spatially co-registered with pixel resolution of 12.5 m.

TABLE IV
TRAINING AND TEST SAMPLES FOR INFORMATION CLASSES IN THE EXPERIMENT ON THE ANDERSON RIVER DATA

Class #	Information Class	Training Size	Test Size
1	Douglas Fir (31-40m)	971	8744
2	Douglas Fir (21-30m)	551	5769
3	Douglas Fir + Other Species(31-40m)	548	4932
4	Douglas Fir + Lodgepole Pine (21-30m)	542	4881
5	Hemlock + Cedar (31-40m)	317	2856
6	Forest Clearings	1260	11340
	Total	4189	38522

There are 19 information classes in the ground reference map provided by CCRS. In the experiments, only the six largest ones were used, as listed in Table IV. Here, training samples were selected uniformly, giving 10% of the total sample size. All other known samples were then used as test samples.

Four statistical methods were used to classify the data: the MED, the Gaussian maximum likelihood method (ML), LOP, and the LOGP. For the LOP and LOGP six data classes (corresponding to the information classes in Table IV) were defined in each data source. The AMSS and SAR data sources were modeled to be Gaussian but the topographic data sources were modeled by Parzen density estimation [30] with Gaussian kernels. The Anderson River Data is very difficult to model [33], and, therefore, few class-conditional source-

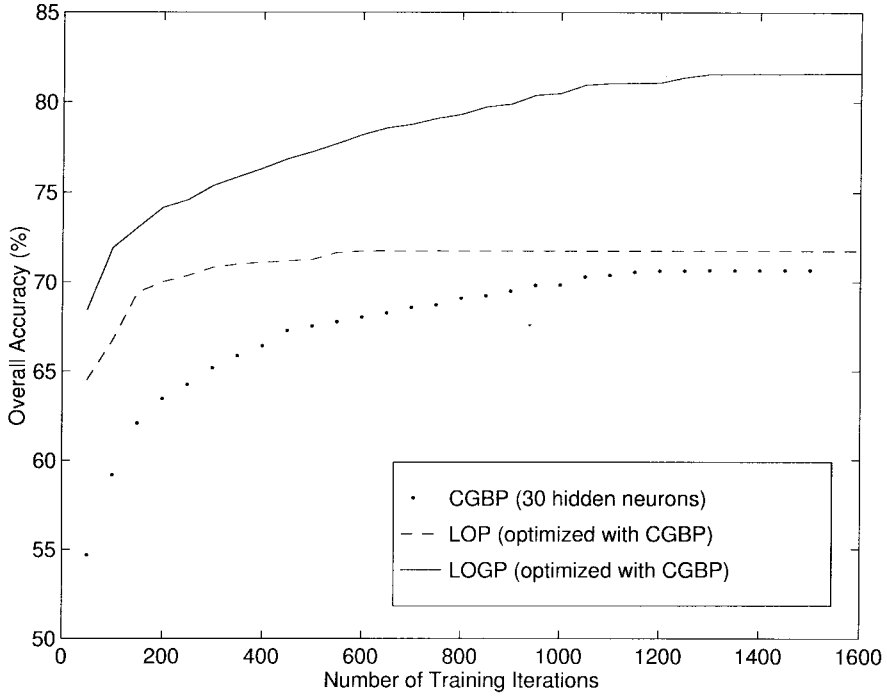

Fig. 6. Anderson river data: Training accuracies as a function of the number of iterations for the methods that utilize the CGBP.

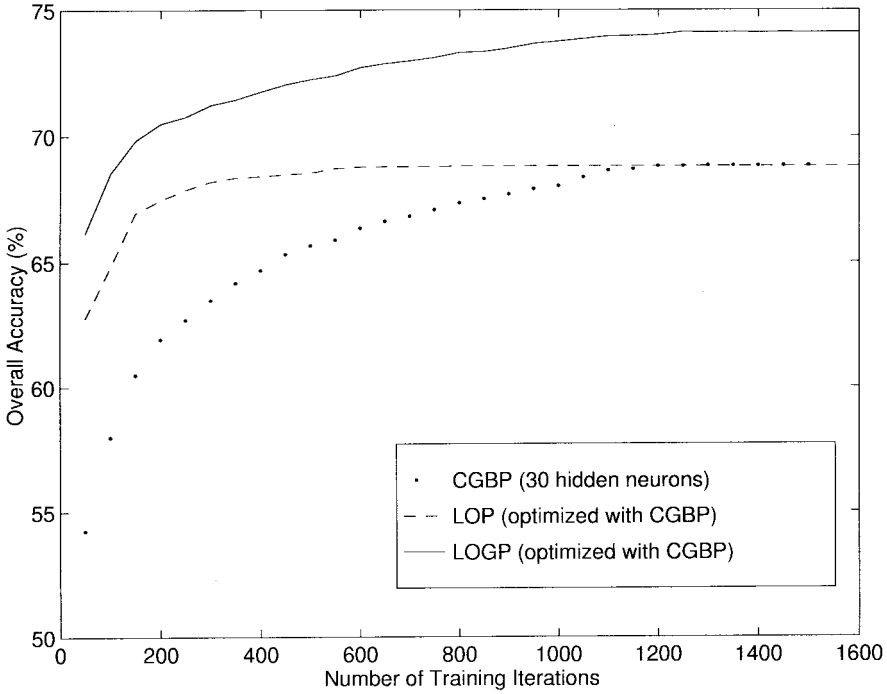

Fig. 7. Anderson river data: Test accuracies as a function of the number of iterations for the methods that utilize the CGBP.

specific probabilities were zero by the probabilistic modeling. Consequently, the veto property of (3) was not a problem for these data.

The same weighting schemes for the LOP and LOGP were used as in the experiment on the Colorado data. For the nonlinear versions, two and three layer CGBP neural networks were utilized with different numbers of hidden neurons (0, 15, 25, 35, and 45 hidden neurons). As in experiment 1, the neural networks were trained six times with different initializations. Then, the average of these six experiments was computed.

As in the experiment on the Colorado data, the CGBP algorithm with two and three layers was trained on the data with different numbers of hidden neurons (0, 15, 30, and 45 hidden neurons). Each version of the CGBP network was trained six times with different initializations and the overall average accuracies were computed in each case (see Figs. 6 and 7).

The overall classification accuracies for the different methods are summarized in Tables V (training) and VI (test). In the tables the average result for the best implementation of the neural network based methods is shown in each case. The

TABLE V
TRAINING ACCURACIES IN PERCENTAGE FOR THE CLASSIFICATION METHODS APPLIED TO THE ANDERSON RIVER DATA SET

Method	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Average Accuracy	Overall Accuracy
MED	40.4	8.9	47.6	67.7	42.3	72.4	46.55	50.51
ML	54.6	31.6	87.8	90.9	81.4	73.3	69.92	68.23
LOP (equal weights)	49.6	0.0	0.0	51.5	0.0	94.9	32.67	47.60
LOP (heuristic weights)	68.2	0.0	0.0	73.1	24.3	89.4	42.50	54.00
LOP (optimal linear weights)	69.8	42.7	81.2	77.5	70.4	78.9	70.07	71.50
LOP (optimized with CGBP)	69.0	45.0	81.3	76.9	85.0	78.4	72.59	71.75
LOGP (equal weights)	68.7	28.1	79.6	78.8	81.7	74.3	68.53	68.75
LOGP (heuristic weights)	68.9	33.2	78.5	79.5	75.7	75.8	68.60	69.42
LOGP (optimal linear weights)	71.9	40.3	79.7	75.1	82.0	79.1	71.35	72.09
LOGP (optimized with CGBP)	81.2	56.0	84.3	88.7	91.7	86.4	81.38	81.62
CGBP (30 hidden neurons)	72.2	34.4	67.2	74.6	79.2	83.1	68.44	70.67
Number of Samples	971	551	548	542	317	1260		4189

TABLE VI
TEST ACCURACIES IN PERCENTAGE FOR THE CLASSIFICATION METHODS APPLIED TO THE ANDERSON RIVER DATA SET

Method	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Average Accuracy	Overall Accuracy
MED	39.7	8.9	48.4	70.2	46.0	71.7	47.48	50.83
ML	50.8	27.7	84.5	81.9	73.8	72.0	64.30	65.12
LOP (equal weights)	49.8	0.0	0.0	50.4	0.0	95.3	32.60	45.76
LOP (heuristic weights)	68.9	0.0	0.0	73.1	20.8	89.3	42.03	53.89
LOP (optimal linear weights)	66.4	34.3	78.5	74.8	72.6	79.5	67.66	68.56
LOP (optimized with CGBP)	67.1	36.7	77.3	75.1	83.4	77.6	69.52	69.15
LOGP (equal weights)	67.9	23.1	77.8	77.5	81.2	73.7	66.85	66.35
LOGP (heuristic weights)	69.0	31.8	75.9	78.6	75.6	75.1	67.60	68.58
LOGP (optimal linear weights)	68.6	32.4	75.2	71.2	81.7	80.1	68.17	68.73
LOGP (optimized with CGBP)	75.4	43.1	76.9	79.5	87.2	82.1	74.02	74.11
CGBP (30 hidden neurons)	71.9	29.3	67.5	73.8	79.3	82.4	67.36	68.83
Number of Samples	8744	5769	4932	4881	2856	11340		38522

following heuristic weights were used for the LOP: AMSS: 1.0, SAR steep mode data: 0.8, SAR shallow mode data: 0.8, Elevation data: 1.0, Slope data: 1.0, and Aspect data: 1.0. The heuristic weights for the LOGP were: AMSS: 1.0, SAR steep mode data: 1.0, SAR Shallow Mode Data: 1.0, Elevation data: 0.0, Slope data: 0.0, and Aspect data: 0.0. The pseudo inverse method of Section III was used as the optimal linear weighting for the consensus theoretic methods.

In Tables V and VI, the conventional classification methods, the MED and ML showed different characteristics. The MED was not acceptable in terms of classification accuracies, but the ML accuracies were relatively good, especially considering that the data are not Gaussian. From the results in Tables V and VI it is clear that the LOGP optimized with a neural network outperformed all other methods in terms of overall and average training and test accuracies. It is noteworthy that the CGBP optimization increased the overall accuracies of the equally weighted LOGP by approximately 12% (training) and 6% (test), and the LOGP with nonlinearly optimized weights outperformed easily the best single stage neural network classifiers both in terms of training and test accuracies. In contrast, the CGBP optimized LOP only gave comparable results to the single stage CGBP with 30 hidden neurons. However, the CGBP optimized LOP improved significantly (between 23% and 30%) on the LOP result with equal weights in terms of average and overall accuracies of training and test data. Also, the best CGBP optimized LOP results were achieved with 0 hidden neurons where the best CGBP optimized LOGP results were reached with 45 hidden neurons. These results are not surprising since the LOP is a linear combination of posterior probabilities but the LOGP is nonlinear. Therefore, a multilayer neural network is needed for the optimization of the LOGP.

V. CONCLUSION

Hybrid statistical/neural network classification has been proposed. The approach is based on using a neural network

as an optimizer for statistical methods which use consensus from several data sources in classification. Two consensus rules were considered: The linear opinion pool (LOP) and the logarithmic opinion pool (LOGP). The LOGP optimized by conjugate gradient backpropagation outperformed all other classification methods used, both in terms of overall and average test accuracies. The results suggest that the discriminative information provided by the equally weighted LOGP can be effective preprocessors for neural networks.

The LOGP is nonlinear and, therefore, a neural network with at least one hidden layer was needed for the optimization of that consensus rule. In fact, the LOGP can be very difficult to optimize because of its veto property. In this respect, the data in experiments 1 and 2 had different characteristics. The Colorado data in experiment 1 had relatively sharp probability density functions. This meant that several probabilities with the value zeros were encountered and, thus, the logarithm of those values approached minus infinity when (3) was used. In contrast, this was not a problem for the LOP. Consequently, in experiment 1 only similar accuracies were achieved for the LOP and LOGP when optimized by the CGBP. On the other hand, the Anderson River data in experiment 2 are extremely difficult to model. Therefore, few 0 class-conditional probabilities were achieved in that experiment since the class-conditional densities were overlapping. For the Anderson River Data, the LOGP became easier to optimize and, consequently, the LOGP optimized by the CGBP easily outperformed the LOP.

Overall, the optimization of the LOP and LOGP worked well. As expected, the accuracies obtained with heuristic weights improved on the accuracies with equal weights, but improvement was again achieved with the optimal linear weights. However, the CGBP optimization was clearly the best in terms of accuracies in both experiments. Based on the results in the paper, the proposed hybrid statistical/neural network consensus theory has potential to be applied successfully for many difficult classification and data fusion problems.

ACKNOWLEDGMENT

This work was done in part while J. A. Benediktsson was a visiting scholar in the School of Electrical and Computer Engineering, Purdue University, W. Lafayette, Indiana. The Colorado data set was originally acquired, preprocessed and loaned to us by Dr. R. Hoffer of Colorado State University. Access to the data set is gratefully acknowledged. The Anderson River SAR/MSS data set was acquired, preprocessed, and loaned by the Canada Centre for Remote Sensing, Department of Energy Mines, and Resources, of the Government of Canada.

REFERENCES

[1] R. C. Luo and M. G. Kay, "Multisensor integration and fusion in intelligent systems," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, pp. 901–931, Sept./Oct. 1989.

[2] J. A. Benediktsson and P. H. Swain, "Consensus theoretic classification methods," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 688–704, July/Aug. 1992.

[3] A. H. Schistad Solberg, A. K. Jain, and T. Taxt, "Multisource classification of remotely sensed data: Fusion of Landsat TM and SAR images," *IEEE Trans. Geosci. Remote Sensing*, vol. 32, pp. 768–778, July 1994.

[4] T. Lee, J. A. Richards, and P. H. Swain, "Probabilistic and evidential approaches for multisource data analysis," *IEEE Trans. Geosci. Remote Sensing*, vol. 25, pp. 283–292, 1987.

[5] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote sensing data," *IEEE Trans. Geosci. Remote Sensing*, vol. 28, no. 4, pp. 540–552, July 1990.

[6] W. Wan and D. Fraser, "A SOM framework for multisource data and contextual analysis," in *Proc. 7th Australasian Remote Sensing Conf.*, 1994, vol. 3, pp. 145–150.

[7] G. A. Carpenter, M. N. Gjaja, S. Gopal, and C. E. Woodcock, "ART neural networks for remote sensing: Vegetation classification from Landsat TM and terrain data," in *Proc. 1996 Int. Geoscience and Remote Sensing Symposium (IGARSS'96)*, Lincoln, NE, May 27–31, 1996, vol. 1, pp. 529–531.

[8] H. Kim and P. H. Swain, "Evidential reasoning approach to multisource-data classification in remote sensing," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, pp. 1257–1265, Aug. 1995.

[9] S.-B. Cho and J. H. Kim, "Multiple network fusion using fuzzy logic," *IEEE Trans. Neural Networks*, vol. 6, pp. 497–501, 1995.

[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*, Lab. Applications of Remote Sensing, Purdue University, West Lafayette, IN, 1975.

[11] A. H. Strahler and N. A. Bryant, "Improving forest cover classification accuracy from landsat by incorporating topographic information," in *Proc. 12th Int. Symp. Remote Sensing of the Environment*, Environmental Inst. Michigan, Apr. 1978, pp. 927–942.

[12] J. Franklin, T. L. Logan, C. E. Woodcock, and A. H. Strahler, "Digital terrain data," *IEEE Trans. Geosci. Remote Sensing*, vol. 25, no. 1, pp. 139–149, 1986.

[13] 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 Sensing*, vol. 9, no. 4, pp. 669–682, 1988.

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

[15] 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, Lab. Applications of Remote Sensing in cooperation with Dept. Forestry and Natural Resources, Purdue University, West Lafayette, IN, 1979.

[16] 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 and 12th Canad. Symp. Remote Sensing*, 1989, vol. 2, pp. 829–832.

[17] A. Srinivasan and J. A. Richards, "Knowledge-based system for multisource classification," *Int. J. Geogr. Inform. Syst.*, vol. 7, pp. 479–500, 1993.

[18] C. E. Woodcock, J. Collins, S. Gopal, V. Jakabhazy, X. Li, S. Macomber, S. Ryherd, Y. Wu, V. J. Harward, J. Levitan, and R. Warbington, "Mapping forest vegetation using Landsat TM imagery and a canopy reflectance model," *Remote Sens. Environ.*, vol. 50, pp. 240–254, 1994.

[19] 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.

[20] A. H. Schistad Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 34, pp. 100–113, 1996.

[21] C. Berenstein, L. N. Kanal, and D. Lavine, "Consensus rules," in *Uncertainty in Artificial Intelligence*, L. N. Kanal and J. F. Lemmer, Eds. New York: North Holland, 1986.

[22] R. F. Bordley, "Studies in mathematical group decision theory," Ph.D. dissertation, Univ. California, Berkeley, 1979.

[23] C. Genest and J. V. Zidek, "Combining probability distributions: A critique and annotated bibliography," *Statist. Sci.*, vol. 1, no. 1, pp. 114–118, 1986.

[24] R. L. Winkler, "Combining probability distributions from dependent information sources," *Manag. Sci.*, vol. 27, pp. 479–488, 1981.

[25] R. A. Jacobs, "Methods for combining experts' probability assessments," *Neural Computat.*, vol. 3, pp. 867–888, 1995.

[26] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[27] D. Graupe, *Time Series Analysis, Identification, and Adaptive Filtering*. Melbourne, FL: Krieger, 1984.
[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[29] P. H. Swain, "Fundamentals of pattern recognition in remote sensing," *Remote Sensing—The Quantitative Approach*, P. H. Swain and S. Davis, Eds. New York: McGraw-Hill, New York, 1978.
[30] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability. New York: Chapman and Hall, 1986.
[31] E. Barnard, "Optimization for training neural nets," *IEEE Trans. Neural Networks*, vol. 3, no. 2, pp. 232–240, Mar. 1992.
[32] D. G. Goodenough, M. Goldberg, G. Plunkett, and J. Zelek, "The CCRS SAR/MSS anderson river data set," *IEEE Trans. Geosci. Remote Sensing*, vol. 25, pp. 360–367, May 1987.
[33] J. A. Benediktsson and P. H. Swain, "Statistical methods and neural network approaches for classification of data from multiple sources," Lab. Applications of Remote Sensing and School Elect. Eng., Purdue University, West Lafayatte, IN, Tech. Rep., TR-EE 90-64, Dec. 1990.

Jon Atli Benediktsson (S'84–M'90) received the C. S. degree in electrical engineering from the University of Iceland, Reykjavik, in 1984, and the M.S.E.E., and the Ph.D. degrees from the School of Electrical Engineering, Purdue University, West Lafayette, IN, in 1987 and 1990, respectively.

He was with the University of Iceland's Laboratory for Information Technology and Signal Processing from 1984 to 1985. He is is currently Associate Professor of Electrical and Computer Engineering, University of Iceland. From 1985 to 1991, he was affiliated with the School of Electrical Engineering and the Laboratory for Applications of Remote Sensing (LARS), Purdue University. His research interests are pattern recognition, neural networks, remote sensing, image processing, and signal processing.

Dr. Benediktsson is the Chairman of the IEEE Geoscience and Remote Sensing Society's Technical Committee on Data Fusion. He received the 1991 Stevan J. Kristof Award as outstanding graduate student in remote sensing from from LARS. He received the 1997 Icelandic Research Council's Young Researcher Award. He is a member of the International Neural Network Society (INNS), the Institute for Nonlinear Analysts (IFNA), and Tau Beta Pi.

Johannes R. Sveinsson (S'86–M'90) received the B.S. degree from the University of Iceland, Reykjavik, and the M.Sc. (Eng.) and the Ph.D. degrees from Queen's University, Kingston, Ont., Canada, all in electrical engineering.

He was a Visiting Research Student at the Department of Electrical Engineering, Imperial College, London, U.K., from 1985 to 1986. From 1981 to 1982, he was with the Laboratory for Information Technology and Signal Processing, University of Iceland. At Queen's University, he held teaching and research assistantship and received the Queen's Graduate Awards. Since November 1991, he has been with the Engineering Research Institute, University of Iceland, as a member of research staff and a Lecturer in the Department of Electrical and Computer Engineering, University of Iceland. His current research interests are in systems theory.

Dr. Sveinsson is a member of SIAM.

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

He is Professor of Electrical and Computer Engineering and Director of Continuing Engineering Education, 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 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 Honorary Visiting Fellow at the University of New South Wales, Sydney, Australia. His research interests have included 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 was Vice Chairman of the Technical Committee on Remote Sensing (TC7) of the International Association for Pattern Recognition, 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. In 1994, he received the Meritorious Achievement Award in Continuing Education from the IEEE Educational Activities board. He is a member of the American Society for Engineering Education (ASEE), the Universities Continuing Education Association, Phi Beta Kappa, Sigma Xi, and Eta Kappa Nu.